

# On Poisson approximations for the Ewens sampling formula when the mutation parameter grows with the sample size

Koji Tsukuda\*

April 25, 2017

## Abstract

The Ewens sampling formula was firstly introduced in the context of population genetics by Warren John Ewens in 1972, and has appeared in a lot of other scientific fields. There are abundant approximation results associated with the Ewens sampling formula especially when one of the parameters, the sample size  $n$  or the mutation parameter  $\theta$  which denotes the scaled mutation rate, tends to infinity while the other is fixed. By contrast, the case that  $\theta$  grows with  $n$  has been considered in a relatively small number of works, although this asymptotic setup is also natural. In this paper, when  $\theta$  grows with  $n$ , we advance the study concerning the asymptotic properties of the total number of alleles and of the counts of components in the allelic partition assuming the Ewens sampling formula, from the viewpoint of Poisson approximations.

---

\*Graduate School of Arts and Sciences, the University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902. mail: ctsukuda@g.ecc.u-tokyo.ac.jp

# 1 Introduction

For a positive integer  $n$ , consider the sequence  $\{C_j^n\}_{j=1}^\infty$  of nonnegative integer-valued random variables satisfying  $\sum_{j=1}^n jC_j^n = n$  and  $C_j^n = 0$  for  $j > n$ . For  $b = 1, \dots, n$ , let us denote  $\mathbf{C}_b^n = (C_1^n, \dots, C_b^n)$  and  $\mathbf{a}_b = (a_1, \dots, a_b)$ . This  $\mathbf{C}_b^n$  denotes the component counts in a random combinatorial structure of size  $n$ . In the context of population genetics, Ewens (1972) introduced what is called *the Ewens sampling formula*

$$P(\mathbf{C}_n^n = \mathbf{a}_n) = \frac{n!}{(\theta)_n} \prod_{j=1}^n \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!} \mathbb{1}\left\{\sum_{j=1}^n ja_j = n\right\} \quad (1.1)$$

as the distribution of the allelic partition in a sample of size  $n$  from the population which follows the stationary distribution of the infinitely-many neutral allele model with scaled mutation rate  $\theta > 0$ , where  $(\theta)_n$  is the rising factorial  $\theta(\theta + 1) \cdots (\theta + n - 1)$ . See for instance Section 2.5 of Feng (2010) for its derivation and basic properties. Hereafter, we consider (1.1) as a model of  $\{C_j^n\}_{j=1}^n$ . The unsigned Stirling number of the first kind  $\bar{s}(n, k)$  ( $k = 1, 2, \dots, n$ ) is the coefficient of  $\theta^k$  in  $(\theta)_n$ , and is in conformity with the number of permutations of  $n$  elements with  $k$  disjoint cycles. Hence, if (1.1) is assumed, the total number  $K_n = \sum_{j=1}^n C_j^n$  of alleles included in the sample, in other words the total number of distinct cycles in a random permutation, follows *the falling factorial distribution* (Watterson, 1974a)

$$P(K_n = k) = \bar{s}(n, k) \frac{\theta^k}{(\theta)_n}. \quad (1.2)$$

In this paper, we will present asymptotic properties, especially Poisson approximations, of  $\mathbf{C}_b^n$  and  $K_n$  when both  $\theta$  and  $n$  increase.

Beyond population genetics domain, the Ewens sampling formula has been widely applied to other fields such as ecology, disclosure risk assessments, nonparametric statistics and so on. In addition, laws of component counts in a lot of random structures are approximated by the Ewens sampling formula. For a general review and an up-to-date review with discussions, we refer the reader to Chapter 41 of Johnson, Kotz and Balakrishnan (1997), whose write-up was provided by S. Tavaré and W.J. Ewens, and to Crane (2016), respectively. For (1.1), (1.2) and related probabilistic models, a lot of works have discussed asymptotic properties under the situations  $n \rightarrow \infty$  with fixed  $\theta$  or  $\theta \rightarrow \infty$  with fixed  $n$ , see for instance Feng (2016). It is natural to consider some relations between the population size and the sample size. Since  $\theta$  is proportional to the population size in the context of population genetics, Feng (2007) and Tsukuda (2017a) discussed the asymptotic behavior of  $K_n$  under the settings that both  $n$  and  $\theta$  simultaneously tend to infinity. Under this asymptotic setting, Feng (2007) established the large deviation principle and Tsukuda (2017a) demonstrated asymptotic properties of the maximum likelihood estimator of  $\theta$ .

Following previous works, we set three major goals. Tsukuda (2017a) extended the asymptotic normality of  $K_n$  as  $n \rightarrow \infty$  with fixed  $\theta$ , which is due to Watterson (1974b), to the situation when both  $\theta$  and  $n$  increase. The first goal of this paper is discussing this result from the viewpoint of Poisson approximations. Moreover, Arratia,

Barbour and Tavaré (1992) showed the Poisson process approximation of  $\mathbf{C}_b^n$  as  $n \rightarrow \infty$  with fixed  $\theta$  when  $b$  is fixed or grows with  $n$ , and Arratia, Stark and Tavaré (1995) established its total variation asymptotics. The second goal is studying corresponding asymptotic results about  $\mathbf{C}_b^n$  when  $\theta$  grows with  $n$ . Furthermore, Hansen (1990) provided a functional central limit theorem for the Ewens sampling formula, and Arratia and Tavaré (1992) gave its elegant proof via the Poisson process approximation. Our third goal is to discuss extensions of this and related weak convergence results.

## 1.1 Notations

Consider sequences  $\{x_n\}_{n=1}^\infty$  and  $\{y_n\}_{n=1}^\infty$ . If  $x_n/y_n \rightarrow 1$ , then we write  $x_n \sim y_n$ . Let  $c < \infty$  be a constant. If  $x_n/y_n \rightarrow 0$  then we write  $x_n = o(y_n)$ , if  $x_n/y_n \rightarrow c$  then we write  $x_n = O(y_n)$ , and if  $x_n/y_n \rightarrow c \neq 0$  then we write  $x_n = \Theta(y_n)$ . Let  $\sum_{j=1}^0 x_j = 0$  and  $\prod_{j=1}^0 x_j = 1$  for any sequence  $\{x_j\}$ , and let  $(x)_0 = 1$  for any value  $x$ . When we consider the limits of  $n$  and  $\theta$  simultaneously, we use the notation  $\lim_{n,\theta}$ .

Let  $[x^k]f(x)$  denote the coefficient of  $x^k$  in the power series expansion of  $f(x)$ . Let  $f^{(i)}(\cdot)$  denote the  $i$ -th derivative of function  $f(\cdot)$ . Let  $\lfloor \cdot \rfloor$  and  $\lceil \cdot \rceil$  denote the floor function and the ceiling function, respectively. Let  $\Gamma(\cdot)$  be the gamma function and  $\psi(\cdot) = (\log \Gamma(\cdot))'$  the digamma function. For real  $x$ ,  $x^+$  denotes the positive part of  $x$ .

The space  $D[0, 1]$  is the set of càdlàg functions on  $[0, 1]$  endowed with the Skorokhod topology. The space  $L^2(0, 1)$  is equivalence classes of real valued functions on  $(0, 1)$  which are square integrable with respect to the Lebesgue measure endowed with the  $L^2$  topology.

The total variation distance between the laws which random vectors  $X$  and  $Y$  follow is denoted by  $d_{TV}(X, Y)$ . The convergence of  $X$  to  $Y$  in probability and the weak convergence of  $X$  to  $Y$  are denoted by  $X \rightarrow^p Y$  and  $X \Rightarrow Y$ , respectively.

## 1.2 Asymptotic settings

Letting  $c$  be a finite constant, we study the following asymptotic settings in this paper:

Case A:  $n/\theta \rightarrow \infty$ ; Case B:  $n/\theta \rightarrow c > 0$ ; Case C:  $n/\theta \rightarrow 0$ ;

Case C1:  $n/\theta \rightarrow 0$  and  $n^2/\theta \rightarrow \infty$ ; Case C2:  $n^2/\theta \rightarrow c > 0$ ; Case C3:  $n^2/\theta \rightarrow 0$ .

This deviation is introduced in Tsukuda (2017a). It should be noted that in Section 4 of Feng (2007), when  $\theta$  does not converge to 0, the relation between  $n$  and  $\theta$  are divided into Cases A, B, C above and  $\theta \rightarrow \infty$  with fixed  $n$ . Moreover, throughout this paper, we assume that  $\theta$  does not decrease as  $n$  increase.

**Remark 1.** In Case C3, it holds that  $K_n - n \rightarrow^p 0$ . Note that when  $\theta = o(1/\log n)$  in which we are not interested since  $\theta \rightarrow 0$ , it holds that  $K_n - 1 \rightarrow^p 0$ . These convergences can be checked through showing the convergence in first mean.

### 1.3 Organization

In Section 2, we review asymptotic results associated with the Ewens sampling formula in the literature which will be discussed in this paper. Before probabilistic result, in Section 3, let us provide some preliminary evaluations for sequences related to the mean of  $K_n$ . Section 4 is devoted to show Poisson approximations for  $K_n$  and  $n - K_n$  in Case A and C, respectively. Section 5 is devoted to discuss independent process approximations for  $\mathbf{C}_b^n$  in a Ewens partition. Section 6 shows the functional central limit theorems for the Ewens sampling formula when  $\theta$  grows with  $n$ . In addition, Appendix includes some lemmas used in proofs.

## 2 Results in the literature

### 2.1 Normal and Poisson approximations for $K_n$

In the combinatorial context, it is worthwhile to know when typical distributions such as Normal, Poisson or other distributions asymptotically appear. See for instance Flajolet and Soria (1990). For the total number  $K_n$  of alleles which follows (1.2), Watterson (1974b) proved the following central limit theorem (CLT for short): For fixed  $\theta > 0$ ,

$$\frac{K_n - \theta \log n}{\sqrt{\theta \log n}} \Rightarrow N(0, 1) \quad (2.1)$$

as  $n \rightarrow \infty$ , where  $N(0, 1)$  is a standard normal variable. A stronger result, the Poisson approximation for  $K_n$ , was stated by Arratia and Tavaré (1992): For fixed  $\theta > 0$ ,

$$d_{TV}(K_n, P_{\mathbb{E}[K_n]}) = \Theta\left(\frac{1}{\log n}\right) \quad (2.2)$$

as  $n \rightarrow \infty$ , where  $P_{\mathbb{E}[K_n]}$  is a Poisson variable with mean  $\mathbb{E}[P_{\mathbb{E}[K_n]}] = \mathbb{E}[K_n]$ . Later, in order to improve the approximation accuracy, Yamato (2013) provided the following CLT which adopts another standardization: For fixed  $\theta > 0$ ,

$$\frac{K_n - \theta(\log n - \psi(\theta))}{\sqrt{\theta(\log n - \psi(\theta))}} \Rightarrow N(0, 1) \quad (2.3)$$

as  $n \rightarrow \infty$ . Moreover, Yamato (2013) showed the approximation for  $K_n$  by a Poisson variable with the approximate mean: For fixed  $\theta > 0$ ,

$$d_{TV}(K_n, P_{\theta(\log n - \psi(\theta))}) = O\left(\frac{1}{\log n}\right) \quad (2.4)$$

as  $n \rightarrow \infty$ , where  $P_{\theta(\log n - \psi(\theta))}$  is a Poisson variable with mean  $\mathbb{E}[P_{\theta(\log n - \psi(\theta))}] = \theta(\log n - \psi(\theta))$ .

When  $\theta$  grows with  $n$ , the standardization should be changed in many cases. Let  $Z_n = (K_n - \mu)/\sigma$ , where  $\mu = \theta \log(1 + n/\theta)$  and  $\sigma^2 = \theta(\log(1 + n/\theta) - n/(n + \theta))$ . Tsukuda (2017a) showed that

$$Z_n \Rightarrow \begin{cases} N(0, 1) & (\text{Case A, B, C1}) \\ (c/2 - P_{c/2})/\sqrt{c/2} & (\text{Case C2}) \\ 0, & (\text{Case C3}) \end{cases} \quad (2.5)$$

where  $c = \lim_{n,\theta} n^2/\theta$  and  $P_{c/2}$  is a Poisson variable with mean  $\mathbb{E}[P_{c/2}] = c/2$ .

**Remark 2.** Professor Shuhei Mano pointed out that the proof of Theorem 2 in Tsukuda (2017a) is incorrect in Case C1. In this remark, let us correct the failure. As it is stated in the right-hand side in the equation (14) of Tsukuda (2017a), it holds that  $\log \mathbb{E}[e^{Z_{nt}}] = -\sigma t + (e^{t/\sigma} - 1)\sigma^2 + A + o(1)$ , where

$$A = \theta \left( \frac{\theta}{n+\theta} - 1 \right) \left\{ \frac{t}{\sigma} - (e^{t/\sigma} - 1) \right\} - \frac{t}{\sigma} \theta e^{t/\sigma} + (\theta e^{t/\sigma} + n) \log \left( 1 + \frac{\theta}{n+\theta} (e^{t/\sigma} - 1) \right).$$

In Case C1, since  $\sigma \sim \sqrt{n^2/2\theta}$ , it holds that  $\theta\{n/(n+\theta)\}^3(-t/\sigma)^3 = O(1/\sqrt{\theta})$ , and hence

$$\begin{aligned} & (\theta e^{t/\sigma} + n) \log \left( 1 + \frac{\theta}{n+\theta} (e^{t/\sigma} - 1) \right) \\ &= (\theta e^{t/\sigma} + n) \log \left( \frac{n + \theta e^{t/\sigma}}{n + \theta} \right) \\ &= (\theta e^{t/\sigma} + n) \log \left( e^{t/\sigma} + \frac{n}{n+\theta} (1 - e^{t/\sigma}) \right) \\ &= (\theta e^{t/\sigma} + n) \left( \frac{t}{\sigma} + \log \left( 1 + \frac{n}{n+\theta} (e^{-t/\sigma} - 1) \right) \right) \\ &= (\theta e^{t/\sigma} + n) \left\{ \frac{t}{\sigma} + \frac{n}{n+\theta} (e^{-t/\sigma} - 1) - \frac{1}{2} \left( \frac{n}{n+\theta} \right)^2 (e^{-t/\sigma} - 1)^2 \right\} + o(1). \end{aligned}$$

We thus have

$$\begin{aligned} A &= -\frac{\theta n}{n+\theta} \frac{t}{\sigma} + \frac{n}{\sigma} t + \frac{n^2}{n+\theta} (e^{-t/\sigma} - 1) - \frac{1}{2} \left( \frac{n}{n+\theta} \right)^2 (\theta e^{t/\sigma} + n) (e^{-t/\sigma} - 1)^2 + o(1) \\ &= \frac{n}{\sigma} t \left( 1 - \frac{\theta}{n+\theta} \right) + n \frac{n}{n+\theta} (e^{-t/\sigma} - 1) - \frac{1}{2} \left( \frac{n}{n+\theta} \right)^2 (\theta e^{t/\sigma} + n) (e^{-t/\sigma} - 1)^2 + o(1) \\ &= n \left( \frac{n}{n+\theta} \right) \left( \frac{t}{\sigma} + e^{-t/\sigma} - 1 \right) - \frac{n^2}{2\theta} \frac{e^{t/\sigma}}{(1+n/\theta)^2} \left( 1 + \frac{ne^{-t/\sigma}}{\theta} \right) (e^{-t/\sigma} - 1)^2 + o(1) \\ &= n \left( \frac{n}{n+\theta} \right) \left( \frac{t}{\sigma} + e^{-t/\sigma} - 1 \right) - \frac{n^2}{2\theta} (e^{-t/\sigma} - 1) (1 - e^{t/\sigma}) \frac{1 + ne^{-t/\sigma}/\theta}{(1+n/\theta)^2} + o(1). \end{aligned} \tag{2.6}$$

By using  $n/(n+\theta) = n/\theta - n^2/\theta^2 + O(n^3/\theta^3)$ , the first term in the right-hand side of (2.6) is

$$n \left( \frac{n}{\theta} - \frac{n^2}{\theta^2} + O\left(\frac{n^3}{\theta^3}\right) \right) \left( \frac{t^2}{2\sigma^2} + O\left(\frac{1}{\sigma^3}\right) \right) = \frac{n^2}{\theta} \frac{t^2}{2\sigma^2} + o(1)$$

The second term in (2.6) is also  $-n^2 t^2 / (2\theta \sigma^2) + o(1)$  because it holds that

$$(e^{-t/\sigma} - 1)(1 - e^{t/\sigma}) = \left( -\frac{t}{\sigma} + \frac{t^2}{2\sigma^2} + o\left(\frac{1}{\sigma^2}\right) \right) \left( -\frac{t}{\sigma} - \frac{t^2}{2\sigma^2} + o\left(\frac{1}{\sigma^2}\right) \right) = \frac{t^2}{\sigma^2} + O\left(\frac{\theta^2}{n^4}\right)$$

and that  $(1 + ne^{-t/\sigma}/\theta)/(1 + n/\theta)^2 = 1 + O(n/\theta)$ . Therefore  $A = o(1)$ , and, consequently,  $\log \mathbb{E}[e^{Z_{nt}}] = -\sigma t + (e^{t/\sigma} - 1)\sigma^2 + o(1) \rightarrow t^2/2$ .

**Remark 3.** As a corollary to the large deviation principle for  $K_n$  when  $\theta \rightarrow \infty$ , Feng (2007) provided the

following weak law of large numbers in Corollary 4.1:

$$\begin{aligned} \frac{K_n}{\theta \log(n/\theta)} &\rightarrow^p 1, \quad (\text{Case A}) \\ \frac{K_n}{n} &\rightarrow^p \begin{cases} \log(1 + \frac{1}{c})^c, & (\text{Case B}) \\ 1, & (\text{Case C}) \end{cases} \end{aligned} \quad (2.7)$$

and  $K_n \rightarrow^p n$  as  $\theta \rightarrow \infty$  with fixed  $n$ . These law of large numbers in Cases A, B and C can be obtained directly from the calculation of  $\mathbb{E}[|K_n/\mathbb{E}[K_n] - 1|^2]$ , see Proposition 2 of Tsukuda (2017a).

## 2.2 Independent process approximations for $\mathbf{C}_b^n$

Consider a sequence  $\{Z_j\}_{j=1}^\infty$  of independent Poisson variables with  $\mathbb{E}[Z_j] = \theta/j$  for  $j = 1, 2, \dots$  and denote  $\mathbf{Z}_b = (Z_1, \dots, Z_b)$  for a positive integer  $b$ . Then, it is well-known that (1.1) can be derived from *the conditioning relation*

$$\mathbb{P}(\mathbf{C}_n^n = \mathbf{a}_n) = \mathbb{P}\left(\mathbf{Z}_n = \mathbf{a}_n \mid \sum_{j=1}^n jZ_j = n\right), \quad (2.8)$$

see for instance Watterson (1974a). It means that the dependence in  $\{\mathbf{C}_j^n\}_{j=1}^\infty$  is given by the condition  $\sum_{j=1}^n jZ_j = n$ . It is of interest to discuss whether the effect of this dependence asymptotically vanishes or not. It was answered by Arratia, Barbour and Tavaré (1992) who showed the small components can be approximated by independent Poisson variables: For any fixed positive integer  $b$ , it holds that

$$(\mathbf{C}_1^n, \dots, \mathbf{C}_b^n) \Rightarrow (\mathbf{Z}_1, \dots, \mathbf{Z}_b) \quad (2.9)$$

as  $n \rightarrow \infty$ . Note that (2.9) is equivalent to  $\lim_{n \rightarrow \infty} d_{TV}(\mathbf{C}_b^n, \mathbf{Z}_b) = 0$  because both  $\mathbf{C}_b^n$  and  $\mathbf{Z}_b$  are discrete.

It is more interesting to consider the case that  $b$  grows with  $n$ . For positive integer  $b$ , let us denote the total variation distance and the distance in the Wasserstein  $\ell^1$  metric between  $\mathbf{C}_b^n = (C_1^n, \dots, C_b^n)$  and  $\mathbf{Z}_b = (Z_1, \dots, Z_b)$  by  $d_b(n)$  and  $d_b^W(n)$ , respectively, that is,

$$\begin{aligned} d_b(n) &= d_{TV}(\mathbf{C}_b^n, \mathbf{Z}_b) = \inf_{\text{couplings}} \mathbb{P}(\mathbf{C}_b^n \neq \mathbf{Z}_b), \\ d_b^W(n) &= \inf_{\text{couplings}} \sum_{j=1}^b \mathbb{E}[|C_j^n - Z_j|]. \end{aligned}$$

For these quantities, it holds that

$$d_b(n) = \inf_{\text{couplings}} \mathbb{P}(\mathbf{C}_b^n \neq \mathbf{Z}_b) = \inf_{\text{couplings}} \mathbb{P}\left(\sum_{j=1}^b |C_j^n - Z_j| \geq 1\right) \leq d_b^W(n).$$

As for the Ewens sampling formula,  $d_b^W(n)$  is a convenient measure of approximations because a concrete construction, *the Feller coupling*, can be given. See Arratia, Barbour and Tavaré (1992, 2016). The Feller

coupling is as follows: Let  $\{\xi_j\}_{j=1}^\infty$  be a sequence of Bernoulli variables with  $P(\xi_j = 1) = p_j = \theta/(\theta + j - 1)$  for any  $j = 1, 2, \dots$ . Then, the Ewens sampling formula (1.1) is given as the joint distribution of

$$C_1^n = \sum_{i=1}^{n-1} \xi_i \xi_{i+1} + \xi_n$$

and

$$C_j^n = \sum_{i=1}^{n-j} \xi_i (1 - \xi_{i+1}) \cdots (1 - \xi_{i+j-1}) \xi_{i+j} + \xi_{n-j+1} (1 - \xi_{n-j+2}) \cdots (1 - \xi_n)$$

for  $j = 2, \dots, n$ . Moreover, define

$$C_j^\infty = \sum_{i=1}^{\infty} \xi_i (1 - \xi_{i+1}) \cdots (1 - \xi_{i+j-1}) \xi_{i+j}$$

for  $j = 1, \dots, n$ , then  $C_j^\infty$  follows the independent Poisson distribution with mean  $E[C_j^\infty] = \theta/j$  for any  $j = 1, 2, \dots$ . That is because the convergences in probability  $\xi_n \xrightarrow{p} 0$  and  $\xi_{n-j+1} (1 - \xi_{n-j+2}) \cdots (1 - \xi_n) \xrightarrow{p} 0$  for any  $j = 2, 3, \dots$  yield that  $C_j^n \Rightarrow C_j^\infty$ , and so (2.9) yields that  $C_j^\infty =^d Z_j$  for any  $j = 1, 2, \dots$ . By using this construction, Arratia, Barbour and Tavaré (1992) proved the Poisson process approximation for  $b$  growing with  $n$ :

$$d_b(n) \rightarrow 0 \Leftrightarrow b = o(n); \quad (2.10)$$

$$d_b(n) \leq \frac{b\theta}{\theta + n} \left( \theta + \frac{n}{\theta + n - b} \right); \quad (2.11)$$

$$d_b^W(n) \leq \frac{b\theta}{\theta + n - b} \left( \theta + \frac{n}{\theta + n} \right); \quad (2.12)$$

$$d_n^W(n) = O(1); \quad (2.13)$$

if  $\theta \geq 1$  then

$$\frac{\theta(\theta - 1)b}{\theta + n - 1} \left\{ 1 - \frac{(\theta - 1)(b + 1)}{4(\theta + n - 1)} \right\} \leq d_b^W(n) \leq \frac{b\theta(\theta + 1)}{\theta + n}. \quad (2.14)$$

Note that (2.11), (2.12) and (2.14) are not asymptotic results. Lower bound results for the total variation distance, which complement (2.11), were given by Arratia, Barbour and Tavaré (1992):  $\liminf_{n \rightarrow \infty} n d_b(n) \geq (b\theta|\theta - 1|/2) \exp\left(-\theta \sum_{j=1}^b 1/j\right)$ ; and by Barbour (1992): if  $\theta \neq 1$  then  $d_b(n) \geq c_3 b/n$  for some  $c_3 = c_3(\theta) > 0$ .

Another compelling result for evaluating  $d_b(n)$  is deriving the leading term of  $d_b(n)$ , which were given by Arratia, Stark and Tavaré (1995) for general logarithmic assemblies. If the Ewens sampling formula is considered, the statement is as follows: If  $b = o(n/\log n)$  then

$$d_b(n) = \frac{|1 - \theta|}{2n} E[|T_{0b} - \theta b|] + o\left(\frac{b}{n}\right), \quad (2.15)$$

where  $T_{0b} = \sum_{j=1}^b j Z_j$ . As it is stated in Corollary 4 of their paper, if  $\theta \neq 1$  and if  $b = o(n/\log n)$  then the leading term of  $d_b(n)$  is given by the first term in the right-hand side of (2.15).

### 2.3 Functional central limit theorems

The results by Arratia, Barbour and Tavaré (1992) provide an elegant way to derive asymptotic properties. Among others, by using (2.13), Arratia and Tavaré (1992) provided an alternative proof of the functional central limit theorem for the Ewens sampling formula which was originally proven by Hansen (1990): The random process

$$X_n^1(\cdot) = \left( \frac{\sum_{i=1}^{\lfloor n^u \rfloor} C_j^n - u\theta \log n}{\sqrt{\theta \log n}} \right)_{0 \leq u \leq 1} \quad (2.16)$$

converges weakly to  $(B(u))_{0 \leq u \leq 1}$  in  $D[0, 1]$  as  $n \rightarrow \infty$ , where  $B(\cdot)$  is a standard Brownian motion. This approach is generalized to broader logarithmic structures. See Arratia, Stark and Tavaré (1995) and Arratia, Barbour and Tavaré (2000). Moreover, by using the Poisson process approximation, Tsukuda (2017b) provided a weighted version in  $L^2(0, 1)$ : Both of the random processes

$$X_n^2(\cdot) = \left( \frac{\sum_{i=1}^{\lfloor n^u \rfloor} C_j^n - \theta \sum_{j=1}^{\lfloor n^u \rfloor} 1/j}{\sqrt{\theta \sum_{j=1}^{\lfloor n^u \rfloor} 1/j}} \right)_{0 < u < 1} \quad (2.17)$$

and

$$X_n^3(\cdot) = \left( \frac{\sum_{i=1}^{\lfloor n^u \rfloor} C_j^n - u\theta \log n}{\sqrt{u\theta \log n}} 1 \left\{ u > \frac{\varepsilon}{\log n} \right\} \right)_{0 < u < 1} \quad (2.18)$$

converge weakly to  $(B(u)/\sqrt{u})_{0 < u < 1}$  in  $L^2(0, 1)$  as  $n \rightarrow \infty$ , where  $\varepsilon$  is a positive constant.

**Remark 4.** In the case that  $\theta = 1$ , the weak convergence of  $X_n^1(\cdot)$  in  $D[0, 1]$  was provided by DeLaurentis and Pittel (1985).

Let  $R_j$  be the  $j$ -th cycle length in a random permutation of  $n$  which has  $K_n$  disjoint cycles, and the *loglength* of  $j$ -th cycle is defined by  $\log_n R_j$ . Consider its empirical distribution function

$$F_n(u) = \frac{\sum_{j=1}^{K_n} 1\{\log_n R_j \leq u\}}{K_n} = \frac{\sum_{j=1}^{\lfloor n^u \rfloor} C_j^n}{K_n}$$

for  $0 \leq u \leq 1$ . Define the random processes

$$X_n^4(\cdot) = \left( \sqrt{\theta \log n} (F_n(u) - u) \right)_{0 \leq u \leq 1}, \quad (2.19)$$

$$X_n^5(\cdot) = \left( \sqrt{\theta \log n} \frac{(F_n(u) - u)}{\sqrt{u(1-u)}} 1 \left\{ \frac{\varepsilon}{\log n} < u < 1 - \frac{\varepsilon}{\log n} \right\} \right)_{0 < u < 1}, \quad (2.20)$$

where  $\varepsilon$  is a positive constant. When  $\theta = 1$ , the weak convergence of  $X_n^4(\cdot)$  to a standard Brownian bridge  $(B^\circ(u))_{0 \leq u \leq 1}$  in  $D[0, 1]$  was shown by DeLaurentis and Pittel (1985), see Notes (2) after Theorem in their paper. Its extension to the Ewens sampling formula and  $L^2$  version are presented as follows, which may have not appeared in the literature.

**Proposition 2.1.** (i) The random process  $X_n^4(\cdot)$  converges weakly to  $(B^\circ(u))_{0 \leq u \leq 1}$  in  $D[0, 1]$  as  $n \rightarrow \infty$ .

(ii) The random process  $X_n^5(\cdot)$  converges weakly to  $(B^\circ(u)/\sqrt{u(1-u)})_{0 \leq u \leq 1}$  in  $L^2(0, 1)$  as  $n \rightarrow \infty$ .



We omit its proof because we will present an extended version in Proposition 6.3. From Proposition 2.1, it follows from the continuous mapping theorem that

$$\sup_{0 \leq u \leq 1} |X_n^4(u)| \Rightarrow \sup_{0 \leq u \leq 1} |B^\circ(u)|, \quad \int_0^1 |X_n^5(u)|^2 du \Rightarrow \int_0^1 \frac{|B^\circ(u)|^2}{u(1-u)} du$$

as  $n \rightarrow \infty$ .

**Remark 5.** *As it is stated in DeLaurentis and Pittel (1985), Proposition 2.1 means that  $F_n(u)$  is nearly  $u$ , which is the distribution function of the standard uniform distribution.*

## 2.4 Auxiliary results

In this subsection, let us set out some auxiliary results concerning Poisson approximations which will be used in the proofs of our statements.

Consider a sequence of independent Bernoulli variables  $\{\xi_j\}_{j=1}^\infty$  and its partial sum  $S_n = \sum_{j=1}^n \xi_j$ , where  $P(\xi_j = 1) = p_j$  for any  $j = 1, 2, \dots$ . Then, by using the Chen–Stein method, Theorems 1 and 2 of Barbour and Hall (1984) gave the sharp bound for the Poisson approximation for a partial sum of Bernoulli variables: For a Poisson variable  $P_\lambda$  with mean  $E[P_\lambda] = \lambda = \sum_{j=1}^n p_j$ , it holds that

$$\frac{1 \wedge \lambda^{-1}}{32} \sum_{j=1}^n p_j^2 \leq d_{TV}(S_n, P_\lambda) \leq \frac{1 - e^{-\lambda}}{\lambda} \sum_{j=1}^n p_j^2. \quad (2.21)$$

Moreover, from a property of the Hellinger integral, a bound for the total variation distance between two Poisson distributions were given in Theorem 2.1 of Yannors (1991): For Poisson variables  $P_{\lambda_1}^1$  and  $P_{\lambda_2}^2$  with respective means  $\lambda_1$  and  $\lambda_2$ , it holds that

$$d_{TV}(P_{\lambda_1}^1, P_{\lambda_2}^2) \leq \min \left( |\sqrt{\lambda_1} - \sqrt{\lambda_2}|, |\lambda_1 - \lambda_2| \right). \quad (2.22)$$

## 3 Preliminary results

Before discussing probabilistic results, let us show asymptotic evaluations on sums of sequences which will be used. Consider two sequences  $\{p_j\}_{j \geq 1}$  and  $\{q_j\}_{j \geq 1}$  given by

$$p_j = \frac{\theta}{\theta + j - 1}, \quad q_j = 1 - p_j = \frac{j - 1}{\theta + j - 1}. \quad (j = 1, 2, \dots, n) \quad (3.1)$$

**Proposition 3.1.** *(i) It holds that*

$$\frac{n}{2(\theta + n)} \leq \sum_{j=1}^n p_j - \theta \log \left( 1 + \frac{n}{\theta} \right) \leq \frac{n}{\theta + n} \quad (3.2)$$

and that

$$0 \leq \sum_{j=1}^n p_j^2 - \frac{n\theta}{n + \theta} \leq 1. \quad (3.3)$$

*Especially, in Case A, it holds that*

$$\sum_{j=1}^n p_j \sim \theta \log \left(1 + \frac{n}{\theta}\right),$$

*and that if  $\theta \rightarrow \infty$  then*

$$\sum_{j=1}^n p_j^2 \sim \theta.$$

*(ii) It holds that*

$$\frac{n(n-1)}{2(\theta+n)} \leq \sum_{j=1}^n q_j \leq \frac{n(n-1)}{2\theta} \quad (3.4)$$

*and that*

$$\frac{n(n-1)(2n-1)}{6(\theta+n)^2} \leq \sum_{j=1}^n q_j^2 \leq \frac{n(n-1)(2n-1)}{6\theta^2}. \quad (3.5)$$

*Especially, in Case C, it holds that*

$$\sum_{j=1}^n q_j \sim \frac{n^2}{2\theta}, \quad \sum_{j=1}^n q_j^2 \sim \frac{n^3}{3\theta^2}.$$

Proof. (i) Since

$$\frac{1}{2} \sum_{j=1}^n \left( \frac{1}{\theta+j-1} - \frac{1}{\theta+j} \right) \leq \sum_{j=1}^n \frac{1}{\theta+j-1} - \int_{\theta}^{\theta+n} \frac{dx}{x} \leq \sum_{j=1}^n \left( \frac{1}{\theta+j-1} - \frac{1}{\theta+j} \right),$$

the result (3.2) holds. Since

$$\int_{\theta}^{\theta+n} \frac{dx}{x^2} \leq \sum_{j=1}^n \frac{1}{(\theta+j-1)^2} \leq \frac{1}{\theta^2} + \int_{\theta}^{\theta+n} \frac{dx}{x^2},$$

the result (3.3) holds.

(ii) From

$$\frac{j-1}{\theta+n} \leq q_j \leq \frac{j-1}{\theta}$$

and from

$$\frac{(j-1)^2}{(\theta+n)^2} \leq q_j^2 \leq \frac{(j-1)^2}{\theta^2},$$

the results (3.4) and (3.5) follow. This completes the proof.

**Proposition 3.2.** *In Case A, it holds that*

$$\sum_{j=1}^n p_j - \theta(\log n - \psi(\theta)) = O\left(\frac{\theta^2}{n}\right). \quad (3.6)$$

Proof. It follows from

$$\sum_{j=1}^n p_j = \theta(\psi(n+\theta) - \psi(\theta)) \quad (3.7)$$

that the left-hand side of (3.6) is  $\theta(\psi(n+\theta) - \log n)$ . Since

$$\psi(n+\theta) = -\gamma - \frac{1}{n+\theta} + (n+\theta) \sum_{j=1}^{\infty} \frac{1}{j(n+\theta+j)}$$

and  $\log n = \sum_{j=1}^n 1/j - \gamma + O(1/n)$  as  $n \rightarrow \infty$ , it holds that

$$\begin{aligned}\psi(n + \theta) - \log n &= -\sum_{j=1}^n \frac{1}{j} + \sum_{j=1}^{\infty} \left( \frac{1}{j} - \frac{1}{n + \theta + j} \right) + O\left(\frac{1}{n}\right) \\ &\leq -\sum_{j=1}^n \frac{1}{j} + \sum_{j=1}^{\infty} \left( \frac{1}{j} - \frac{1}{n + \lfloor \theta \rfloor + 1 + j} \right) + O\left(\frac{1}{n}\right) \\ &= \sum_{j=n+1}^{n+\lfloor \theta \rfloor} \frac{1}{j} + O\left(\frac{1}{n}\right).\end{aligned}$$

In Case A, the first term in the right-hand side is  $O(\theta/n)$ . This completes the proof.

**Remark 6.** As it is stated in (2.3), Yamato (2013) discussed the asymptotic normality of  $K_n$  standardized by  $\theta(\log n - \psi(\theta))$ , which means that  $\psi(n + \theta)$  is approximated by  $\log n$  from (3.7). If  $\theta^3/(n^2 \log(n/\theta)) \rightarrow \infty$ , the bound in (3.6) is meaningless to discuss CLT. On the other hand, if  $\theta^2/n \rightarrow 0$  the centering by  $\theta(\log n - \psi(\theta))$  is better than centering by  $\theta \log(1 + n/\theta)$ , which was used in Corollary 2 of Tsukuda (2017a), because  $\sum_{j=1}^n p_j - \theta \log(1 + n/\theta) = \Theta(n/(n + \theta))$ .

**Proposition 3.3.** In Case C, it holds that

$$\sum_{j=1}^n p_j - \sum_{j=1}^n \frac{\theta}{j} \left( \frac{n}{\theta} \right)^j = O\left( \frac{n^2}{\theta} \right).$$

Proof. The triangle inequality yields that

$$\begin{aligned}&\left| \sum_{j=1}^n \left\{ p_j - \frac{\theta}{j} \left( \frac{n}{\theta} \right)^j \right\} \right| \\ &\leq \left| \sum_{j=1}^n p_j - \theta \log \left( 1 + \frac{n}{\theta} \right) \right| + \left| \theta \log \left( 1 + \frac{n}{\theta} \right) - \log \left( 1 - \frac{n}{\theta} \right)^{-\theta} \right| \\ &\quad + \left| \log \left( 1 - \frac{n}{\theta} \right)^{-\theta} - \sum_{j=1}^n \frac{\theta}{j} \left( \frac{n}{\theta} \right)^j \right|.\end{aligned}$$

The first term is  $O(n/(n + \theta)) = O(n/\theta)$ , the second term is  $\theta \log(1 - n^2/\theta^2) = O(n^2/\theta)$ , and from  $\log(1 - x)^{-1} = x + x^2/2 + \dots$  as  $x \rightarrow 0$  the third term is

$$\sum_{j=n+1}^{\infty} \frac{\theta}{j} \left( \frac{n}{\theta} \right)^j \leq \frac{\theta}{n} \sum_{j=n+1}^{\infty} \left( \frac{n}{\theta} \right)^j = \frac{\theta}{n} \frac{(n/\theta)^{n+1}}{1 - n/\theta} = O\left( \left( \frac{n}{\theta} \right)^n \right).$$

This completes the proof.

## 4 Poisson approximations for the total number of alleles

Introduce two Poisson variables  $Z^A$  and  $Z^C$  whose means are given by  $\lambda_A = \mathbb{E}[K_n] = \sum_{j=1}^n p_j$  and  $\lambda_C = n - \mathbb{E}[K_n] = \sum_{j=1}^n q_j$ , respectively, where  $K_n$  follows (1.2) and where  $\{p_j\}_{j=1}^{\infty}$  and  $\{q_j\}_{j=1}^{\infty}$  are given in (3.1).

Poisson approximations corresponding to (2.2) are given in the following proposition.

**Proposition 4.1.** (i) In Case A,

$$d_{TV}(K_n, Z^A) \leq \frac{n\theta + n + \theta}{\theta(n + \theta) \log(1 + n/\theta) + n/2}, \quad (4.1)$$

and

$$d_{TV}(K_n, Z^A) = \Theta\left(\frac{1}{\log(n/\theta)}\right).$$

(ii) In Case C,

$$d_{TV}(n - K_n, Z^C) \leq \frac{2n(n + \theta)}{3\theta^2}(1 - e^{-n^2/2\theta}), \quad (4.2)$$

and

$$d_{TV}(n - K_n, Z^C) = \Theta\left(\frac{n}{\theta} \left(1 \wedge \frac{n^2}{\theta}\right)\right).$$

Proof. Let  $\{\xi_j\}_{j=1}^\infty$  and  $\{\zeta_j\}_{j=1}^\infty$  be sequences of Bernoulli variables with respective parameters  $P(\xi_j = 1) = p_j$  and  $P(\zeta_j = 1) = q_j$  for  $j = 1, 2, \dots$ . Then, it holds that  $K_n =^d \sum_{i=1}^n \xi_i$  and that  $n - K_n =^d \sum_{i=1}^n \zeta_i$ . To prove the desired results, we will use (2.21) and Proposition 3.1.

(i) The result (4.1) follows from

$$d_{TV}(K_n, Z^A) \leq \frac{\sum_{j=1}^n p_j^2}{\sum_{j=1}^n p_j} \leq \frac{n\theta/(n + \theta) + 1}{\theta \log(1 + n/\theta) + n/\{2(\theta + n)\}}.$$

Since  $\sum_{j=1}^n p_j \rightarrow \infty$ , it holds that

$$d_{TV}(K_n, Z^A) \geq \frac{\sum_{j=1}^n p_j^2}{32 \sum_{j=1}^n p_j} \geq \frac{n\theta/(n + \theta)}{32\theta \log(1 + n/\theta)}$$

for enough large  $n$ . Two displays above imply  $d_{TV}(K_n, Z^A) = \Theta(1/\log(n/\theta))$ .

(ii) The result (4.2) follows from

$$\begin{aligned} d_{TV}(n - K_n, Z^C) &\leq (1 - e^{-\sum_{j=1}^n q_j}) \frac{\sum_{j=1}^n q_j^2}{\sum_{j=1}^n q_j} \\ &\leq (1 - e^{-n^2/2\theta}) \frac{n(n-1)(2n-1)/(6\theta^2)}{n(n-1)/\{2(\theta + n)\}}. \end{aligned}$$

In Case C1, since  $\sum_{j=1}^n q_j \rightarrow \infty$ , it holds that

$$d_{TV}(n - K_n, Z^C) \geq \frac{\sum_{j=1}^n q_j^2}{32 \sum_{j=1}^n q_j} \geq \frac{n(n-1)(2n-1)/\{6(\theta + n)^2\}}{32n(n-1)/2\theta}$$

for enough large  $n$ . Two displays above imply the  $d_{TV}(n - K_n, Z^C) = \Theta(n/\theta)$ . In Case C2, since  $1 - e^{-\sum_{j=1}^n q_j} \leq 1$  and since  $(1/\sum_{j=1}^n q_j)$  is bounded by some constant for enough large  $n$ , the same evaluation provides  $d_{TV}(n - K_n, Z^C) = \Theta(n/\theta)$ . In Case C3, since  $\sum_{j=1}^n q_j \rightarrow 0$ , it holds that  $1 - e^{-\sum_{j=1}^n q_j} \sim n^2/(2\theta)$  and that

$$d_{TV}(n - K_n, Z^C) \geq \frac{1}{32} \sum_{j=1}^n q_j^2 \geq \frac{n(n-1)(2n-1)}{192(\theta + n)^2}$$

for enough large  $n$ . We thus have  $d_{TV}(n - K_n, Z^C) = \Theta(n^3/\theta^2)$ . This completes the proof.

**Remark 7.** From asymptotic properties of the Poisson distribution and Proposition 4.1, the result of (2.5) in Cases A and C can be derived.

In Proposition 4.1, we have considered Poisson variables with rigorous means  $\mathbb{E}[K_n]$  and  $n - \mathbb{E}[K_n]$ . Next, let us discuss centerings by approximate means presented by Yamato (2013) and Tsukuda (2017a) from the viewpoint of Poisson approximation. Introduce three Poisson variables  $Y^A, Y^a$  and  $Y^C$  whose means are given by  $\mu_A = \mathbb{E}[Y^A] = \theta \log(1 + n/\theta)$ ,  $\mu_a = \mathbb{E}[Y^a] = \theta(\log n - \psi(\theta))$  and  $\mu_C = \mathbb{E}[Y^C] = n - \theta \log(1 + n/\theta)$ , respectively.

**Lemma 4.2.** (i) In Case A, it holds that

$$d_{TV}(Z^A, Y^A) = O\left(\frac{1}{\sqrt{\theta \log(n/\theta)}}\right) \quad (4.3)$$

and that

$$d_{TV}(Z^A, Y^a) = O\left(\frac{\theta^{3/2}}{n\sqrt{\log(n/\theta)}}\right). \quad (4.4)$$

(ii) In Case C, it holds that

$$d_{TV}(Z^C, Y^C) = O\left(\frac{1}{\sqrt{\theta}} \left(1 \wedge \frac{n}{\sqrt{\theta}}\right)\right). \quad (4.5)$$

Proof. We will use (2.22). (i) First we see (4.3). Since  $\lambda_A$  and  $\mu_A$  tend to infinity in Case A,  $|\sqrt{\lambda_A} - \sqrt{\mu_A}| \leq |\lambda_A - \mu_A|$  for enough large  $n$ . Moreover, by using Proposition 3.1, it holds that

$$|\sqrt{\lambda_A} - \sqrt{\mu_A}| = \frac{|\lambda_A - \mu_A|}{\sqrt{\lambda_A} + \sqrt{\mu_A}} = \frac{|\sum_{j=1}^n p_j - \theta \log(1 + n/\theta)|}{\sqrt{\sum_{j=1}^n p_j} + \sqrt{\theta \log(1 + n/\theta)}} = O\left(\frac{n/(n + \theta)}{\sqrt{\theta \log(1 + n/\theta)}}\right),$$

and hence (4.3).

Next we see (4.4). By using Propositions 3.1 and 3.2, it holds that

$$|\sqrt{\lambda_A} - \sqrt{\mu_a}| = \frac{|\lambda_A - \mu_a|}{\sqrt{\lambda_A} + \sqrt{\mu_a}} = \frac{|\sum_{j=1}^n p_j - \theta(\log n - \psi(\theta))|}{\sqrt{\sum_{j=1}^n p_j} + \sqrt{\theta(\log n - \psi(\theta))}} = O\left(\frac{\theta^2/n}{\sqrt{\theta \log(1 + n/\theta)}}\right),$$

and hence (4.4).

(ii) First consider Case C1. Since  $\lambda_C$  and  $\mu_C$  tend to infinity in Case C1,  $|\sqrt{\lambda_C} - \sqrt{\mu_C}| \leq |\lambda_C - \mu_C|$  for enough large  $n$ . By using Proposition 3.1, it holds that

$$|\sqrt{\lambda_C} - \sqrt{\mu_C}| = \frac{|\lambda_C - \mu_C|}{\sqrt{\lambda_C} + \sqrt{\mu_C}} = \frac{|\sum_{j=1}^n q_j - (n - \theta \log(1 + n/\theta))|}{\sqrt{\sum_{j=1}^n q_j} + \sqrt{n - \theta \log(1 + n/\theta)}} = O\left(\frac{n/(\theta + n)}{\sqrt{n^2/\theta}}\right) = O\left(\frac{1}{\sqrt{\theta}}\right),$$

and hence (4.5) holds as  $d_{TV}(Z^C, Y^C) = O(1/\sqrt{\theta})$ .

Next consider Case C2. The magnitude relationship of  $|\sqrt{\lambda_C} - \sqrt{\mu_C}|$  and  $|\lambda_C - \mu_C|$  is not determined, but they have the same bound  $O(1/\sqrt{\theta})$  because  $1/\sqrt{\theta} = \Theta(1/n)$ . Hence (4.5) holds as  $d_{TV}(Z^C, Y^C) = O(1/\sqrt{\theta})$ .

Finally, consider Case C3. Since  $\lambda_C$  and  $\mu_C$  tend to 0,  $|\sqrt{\lambda_C} - \sqrt{\mu_C}| \geq |\lambda_C - \mu_C|$  for enough large  $n$ . By using Proposition 3.1, it holds that

$$|\lambda_C - \mu_C| = \left| \sum_{j=1}^n q_j - \left(n - \theta \log\left(1 + \frac{n}{\theta}\right)\right) \right| = O\left(\frac{n}{\theta}\right),$$

and hence (4.5) holds as  $d_{TV}(Z^C, Y^C) = O(n/\theta)$ . This completes the proof.

From what has already been proven, the triangle inequality yields the following Poisson approximations corresponding to (2.4).

**Proposition 4.3.** (i) In Case A, if  $(\log(n/\theta))/\theta \rightarrow \infty$  then

$$d_{TV}(K_n, Y^A) = O\left(\frac{1}{\sqrt{\theta \log(n/\theta)}}\right),$$

and if  $(\log(n/\theta))/\theta = O(1)$  then

$$d_{TV}(K_n, Y^A) = O\left(\frac{1}{\log(n/\theta)}\right).$$

Moreover, in Case A, if  $(\theta^3 \log(n/\theta))/n^2 = O(1)$  then

$$d_{TV}(K_n, Y^a) = O\left(\frac{1}{\log(n/\theta)}\right),$$

and if  $(\theta^3 \log(n/\theta))/n^2 \rightarrow \infty$  and  $\theta^3/(n^2 \log(n/\theta)) = O(1)$  then

$$d_{TV}(K_n, Y^a) = O\left(\frac{\theta^{3/2}}{n\sqrt{\log(n/\theta)}}\right).$$

(ii) In Case C, it holds that

$$d_{TV}(n - K_n, Y^C) = O\left(\frac{n}{\theta}\right).$$

## 5 On independent process approximation of component counts

### 5.1 Case A

First we see the asymptotic independence of small components  $\mathbf{C}_b^n = (C_1^n, \dots, C_b^n)$  in Case A when  $\theta \geq 1$  for some  $n$ , recalling that we assume that  $\theta$  does not decrease as  $n$  increase. We will not discuss the other case,  $\theta < 1$  for all  $n$ , because we are interested in large  $\theta$ .

Consider  $\{Z_j\}_{j=1}^\infty$  defined in Subsection 2.2. Let us denote  $\mathbf{Z}_b = (Z_1, \dots, Z_b)$  for  $b = 1, 2, \dots, n$ , and  $T_{lm} = \sum_{j=l+1}^m jZ_j$  for  $l = 0, 1, \dots, n-1$  and  $m = l+1, \dots, n$ . It follows from the conditioning relation (2.8) that

$$\mathbf{P}(\mathbf{C}_b^n = \mathbf{a}_b) = \mathbf{P}(\mathbf{Z}_b = \mathbf{a}_b) \frac{\mathbf{P}(T_{bn} = n - a)}{\mathbf{P}(T_{0n} = n)}, \quad (5.1)$$

where  $\mathbf{a}_b = (a_1, \dots, a_b)$  and  $a = \sum_{j=1}^b ja_j$ .

**Proposition 5.1.** In Case A, if  $\theta \geq 1$  for some  $n$  and if  $\theta^2/n \rightarrow 0$ , then  $\mathbf{P}(\mathbf{C}_b^n = \mathbf{a}_b) \sim \mathbf{P}(\mathbf{Z}_b = \mathbf{a}_b)$  for any  $\mathbf{a}_b$  with any fixed positive integer  $b$ .

Proof. From (5.1), in order to prove the desired result, it suffices to show that

$$\frac{\mathbb{P}(T_{bn} = n - a)}{\mathbb{P}(T_{0n} = n)} \rightarrow 1.$$

We first calculate  $g_{n-a} = \exp(\theta \sum_{j=b+1}^n 1/j) \mathbb{P}(T_{bn} = n - a)$ . Letting  $f(x) = \exp(-\theta \sum_{j=1}^b x^j/j)$ , we have

$$g_{n-a} = [x^{n-a}](1-x)^{-\theta} f(x),$$

see equation (5) of Arratia, Barbour and Tavaré (1992).

Let  $n$  be a positive integer such that  $\theta \geq 1$ . It holds that

$$\begin{aligned} & [x^{n-a}](1-x)^{-\theta} f(x) \\ = & [x^{n-a}](1-x)^{-\theta} \left\{ \sum_{k=0}^{\lceil \theta \rceil - 1} \frac{f^{(k)}(1)}{k!} (-1)^k (1-x)^k + \sum_{k=\lceil \theta \rceil}^{\infty} \frac{f^{(k)}(1)}{k!} (-1)^k (1-x)^k \right\} \\ = & [x^{n-a}](1-x)^{-\theta} \left\{ \sum_{k=0}^{\lceil \theta \rceil - 1} \frac{f^{(k)}(1)}{k!} (-1)^k (1-x)^k \right\} + [x^{n-a}]h(x), \end{aligned} \quad (5.2)$$

where

$$h(x) = (1-x)^{-\theta} \sum_{k=\lceil \theta \rceil}^{\infty} \frac{f^{(k)}(1)}{k!} (-1)^k (1-x)^k.$$

Since the right-hand side of (5.2) is

$$\frac{f(1)(\theta)_{n-a}}{(n-a)!} \left\{ 1 + \sum_{k=1}^{\lceil \theta \rceil - 1} \frac{(-1)^k}{k!} \frac{f^{(k)}(1)}{f(1)} \frac{(\theta - k)_{n-a}}{(\theta)_{n-a}} \right\} + [x^{n-a}]h(x), \quad (5.3)$$

the first term and the second term will be evaluated in Lemma 5.2 and Lemma 5.3, respectively. From Lemma 5.2, the elements in the bracket of the first term is  $1 + O(\theta^2/n)$ . Next we see  $[x^{n-a}]h(x)$ . It follows from Lemma 5.3 that

$$|[x^{n-a}]h(x)| \leq \frac{1}{r_1^n} \left( \frac{be^{(r_2-1)^b}}{r_2} \right)^\theta \frac{1}{r_2 - 1 - r_1} \{r_1^a(1+r_1)r_2\},$$

where  $r_1 = 1 + c_1 r$  and  $r_2 = 2 + c_2 r$  with constants  $c_1, c_2$  such that  $1 < c_1 < c_2$ . By letting  $r$  be a positive constant, the right-hand side is  $o(1/n^k)$  for any positive  $k$  since  $b$  is fixed and since  $\theta^2/n \rightarrow 0$ . Thence  $[x^{n-a}]h(x) = o(1/n)$ .

Now we have

$$g_{n-a} = f(1) \frac{(\theta)_{n-a}}{(n-a)!} \left( 1 + O\left(\frac{\theta^2}{n}\right) \right) + o\left(\frac{1}{n}\right) \quad (5.4)$$

and, as a result,

$$\begin{aligned} & \mathbb{P}(T_{bn} = n - a) \\ = & \exp\left(-\theta \sum_{j=1}^n \frac{1}{j}\right) \frac{(\theta)_{n-a}}{(n-a)!} \left( 1 + O\left(\frac{\theta^2}{n}\right) \right) + \exp\left(-\theta \sum_{j=b+1}^n \frac{1}{j}\right) o\left(\frac{1}{n}\right). \end{aligned}$$

On the other hand,

$$\mathbb{P}(T_{0n} = n) = \exp \left( -\theta \sum_{j=1}^n \frac{1}{j} \right) [x^n](1-x)^{-\theta} = \exp \left( -\theta \sum_{j=1}^n \frac{1}{j} \right) \frac{(\theta)_n}{n!}. \quad (5.5)$$

If  $\theta \rightarrow c < \infty$ ,  $(\theta)_n/n! \sim n^{\theta-1}/\Gamma(\theta)$  and so

$$\exp \left( \theta \sum_{j=1}^b \frac{1}{j} \right) \frac{n!}{(\theta)_n} \leq \exp(\theta \log b) \frac{e^\theta n!}{(\theta)_n} \sim n\Gamma(\theta) \left( \frac{be}{n} \right)^\theta = o(n). \quad (5.6)$$

If  $\theta \rightarrow \infty$ , Lemma A.1 and the Stirling formula yield that

$$\frac{(\theta)_n}{n!} \sim \frac{n^{\theta-1}}{\Gamma(\theta)} \sim \frac{n^{\theta-1}\theta^{1/2}e^\theta}{\sqrt{2\pi}\theta^\theta},$$

and hence

$$\exp \left( \theta \sum_{j=1}^b \frac{1}{j} \right) \frac{n!}{(\theta)_n} \sim \left( \frac{\theta e^{(\sum_{j=1}^b \frac{1}{j} - 1)}}{n} \right)^\theta \frac{\sqrt{2\pi}n}{\theta^{1/2}} \leq \sqrt{2\pi} \frac{n}{\theta^{1/2}} \left( \frac{\theta b}{n} \right)^\theta = o(n). \quad (5.7)$$

From what has already been proven, we obtain

$$\begin{aligned} \frac{\mathbb{P}(T_{bn} = n-a)}{\mathbb{P}(T_{0n} = n)} &= \frac{n!}{(\theta)_n} \frac{(\theta)_{n-a}}{(n-a)!} \left( 1 + O\left(\frac{\theta^2}{n}\right) \right) + o(n)o\left(\frac{1}{n}\right) \\ &\sim \frac{\theta^\theta}{n^{\theta-1}\theta^{1/2}e^\theta} \frac{(n-a)^{\theta-1}\theta^{1/2}e^\theta}{\theta^\theta} \rightarrow 1. \end{aligned}$$

This completes the proof.

In the following lemma, we see the first term of (5.3).

**Lemma 5.2.** *Let  $f(x) = \exp(-\theta \sum_{j=1}^b x^j/j)$ . For  $\theta > 1$ , for  $k = 1, \dots, [\theta] - 1$  and for any positive integers  $a < n$  and  $b$ , it holds that*

$$\frac{1}{k!} \left| \frac{f^{(k)}(1)}{f(1)} \right| \frac{(\theta - k)_{n-a}}{(\theta)_{n-a}} \leq \left( \frac{b\theta^2}{n-a} \right)^k. \quad (5.8)$$

Moreover, in Case A, if  $a = o(n)$ ,  $b = o(n/\theta^2)$ ,  $\theta \geq 1$  for some  $n$  and  $\theta^2/n \rightarrow 0$ , then

$$\sum_{k=0}^{[\theta]-1} \frac{1}{k!} \left| \frac{f^{(k)}(1)}{f(1)} \right| \frac{(\theta - k)_{n-a}}{(\theta)_{n-a}} = 1 + \frac{b\theta(\theta-1)}{\theta + n - a} + O\left(\frac{b^2\theta^4}{n^2}\right). \quad (5.9)$$

Proof. Let  $g(x)$  be  $-\theta \sum_{j=1}^b x^j/j$ . It holds that

$$g^{(i)}(x) = -\theta \sum_{j=i}^b \frac{j \cdots (j-i+1)}{j} x^{j-i} = -\theta \sum_{j=1}^{b-i+1} (j)_{i-1} x^{j-1}$$

for  $1 \leq i \leq b$ . Thus, for  $1 \leq i \leq b$ ,

$$0 \geq g^{(i)}(1) = -\theta \sum_{j=1}^{b-i+1} (j)_{i-1} \geq -\theta(b-i+1)_{i-1}(b-i+1) \geq -\theta b^i.$$



For  $i > b$ ,  $g^{(i)}(1) = 0 \geq -\theta b^i$ . The Faà di Bruno formula yields that

$$f^{(k)}(x) = \exp(g(x)) \sum_{j=1}^k B_{k,j}((g^{(1)}(x), \dots, g^{(k-j+1)}(x))),$$

where  $B_{k,j}(\cdot)$  is the partial Bell polynomial, so

$$\frac{f^{(k)}(1)}{f(1)} = \sum_{j=1}^k B_{k,j}((g^{(1)}(1), \dots, g^{(k-j+1)}(1)))$$

for any  $k = 1, 2, \dots$ . By using the triangle inequality,

$$\begin{aligned} |B_{k,j}((g^{(1)}(1), \dots, g^{(k-j+1)}(1)))| &\leq n! \sum_{\{s: \sum s_i = j, \sum i s_i = k\}} \prod_{i=1}^n \left( \frac{|g^{(i)}(1)|}{i!} \right)^{s_i} \frac{1}{s_i!} \\ &\leq \theta^j b^k \mathcal{S}(k, j), \end{aligned}$$

where  $\mathcal{S}(k, j)$  is the Stirling number of the second kind. The above two displays and the triangle inequality imply that

$$\begin{aligned} \left| \frac{f^{(k)}(1)}{f(1)} \right| &\leq \sum_{j=1}^k |B_{k,j}((g^{(1)}(1), \dots, g^{(k-j+1)}(1)))| \\ &\leq b^k \sum_{j=1}^k \theta^j \mathcal{S}(k, j) \\ &\leq b^k \sum_{j=1}^k \theta^j \bar{s}(k, j) = b^k (\theta)_k. \end{aligned}$$

For  $k \leq \lceil \theta \rceil - 1$ , the Stirling formula yields that

$$(\theta)_k \frac{(\theta - k)_{n-a}}{(\theta)_{n-a}} = \frac{\Gamma(\theta + k) \Gamma(\theta - k + n - a) \Gamma(\theta)}{\Gamma(\theta) \Gamma(\theta - k) \Gamma(\theta + n - a)} \leq \frac{\theta^{2k}}{(n - a)^k},$$

where we have used  $\Gamma(\theta + k)/\Gamma(\theta - k) = (\theta - k)(\theta^2 - (k - 1)^2) \cdots (\theta^2 - 1^2)\theta \leq \theta^{2k}$  and  $\Gamma(\theta - k + n - a)/\Gamma(\theta + n - a) = 1/((\theta - 1 + n - a) \cdots (\theta - k + n - a)) \leq 1/(n - a)^k$ . We thus have

$$\left| \frac{f^{(k)}(1)}{f(1)} \right| \frac{(\theta - k)_{n-a}}{(\theta)_{n-a}} \leq \left( \frac{b\theta^2}{n - a} \right)^k$$

for  $k \leq \lceil \theta \rceil - 1$ , which is (5.8).

Next we prove (5.9). If  $\theta \leq 1$  for all  $n$ , the result is obvious because the left-hand side of (5.9) is 1. Otherwise, by letting  $n$  be an positive integer such that  $\theta > 1$ , the desired result follows from

$$\frac{1}{1!} \frac{f^{(1)}(1)}{f(1)} \frac{(\theta - 1)_{n-a}}{(\theta)_{n-a}} = b\theta \frac{\theta - 1}{\theta + n - a}$$

and from

$$\sum_{k=2}^{\lceil \theta \rceil - 1} \frac{1}{k!} \left| \frac{f^{(k)}(1)}{f(1)} \right| \frac{(\theta - k)_{n-a}}{(\theta)_{n-a}} \leq \sum_{k=2}^{\lceil \theta \rceil - 1} \left( \frac{b\theta^2}{n - a} \right)^k = O \left( \left( \frac{b\theta^2}{n} \right)^2 \right). \quad (5.10)$$

This completes the proof.

The following lemma is used to evaluate  $[x^{n-a}]h(x)$  in (5.3).

**Lemma 5.3.** Let  $f(x) = \exp(-\theta \sum_{j=1}^b x^j/j)$  and let

$$h(x) = (1-x)^{-\theta} \sum_{k=\lceil \theta \rceil}^{\infty} \frac{f^{(k)}(1)}{k!} (-1)^k (1-x)^k.$$

Then, for any positive integers  $a < n$  and  $b$ , it holds that

$$|[x^{n-a}]h(x)| \leq \frac{1}{r_1^n} \left( \frac{be^{(r_2-1)^b}}{r_2} \right)^{\theta} \frac{1}{r_2 - 1 - r_1} \{r_1^a (1+r_1)r_2\},$$

where  $r_1 = 1 + c_1 r$ ,  $r_2 = 2 + c_2 r$ ,  $1 < c_1 < c_2$ , and  $r$  is an arbitrary positive constant.

Proof. Consider a complex variable  $\mathbf{z} \in \mathbb{C}$ . Since  $h(\mathbf{z})$  and  $f(\mathbf{z})$  are analytic in  $\mathbb{C}$ , by using the Cauchy inequality for coefficients, it holds that

$$\begin{aligned} \sup_{|\mathbf{z}|=r_1} |h(\mathbf{z})| &\leq \sup_{|\mathbf{z}|=r_1} \sum_{k=\lceil \theta \rceil}^{\infty} \left| \frac{f^{(k)}(1)}{k!} \right| |(1-\mathbf{z})^{k-\theta}| \\ &\leq \sum_{k=\lceil \theta \rceil}^{\infty} \left| \frac{f^{(k)}(1)}{k!} \right| \sup_{|\mathbf{z}|=r_1} |(1-\mathbf{z})^{k-\theta}| \\ &\leq \sum_{k=\lceil \theta \rceil}^{\infty} \frac{\sup_{|1-\mathbf{z}|=r_2} |f(\mathbf{z})|}{r_2^k} (1+r_1)^{k-\theta} \\ &= \frac{\sup_{|1-\mathbf{z}|=r_2} |f(\mathbf{z})|}{(1+r_1)^{\theta}} \sum_{k=\lceil \theta \rceil}^{\infty} \left( \frac{1+r_1}{r_2} \right)^k \\ &= \frac{\sup_{|1-\mathbf{z}|=r_2} |f(\mathbf{z})|}{(1+r_1)^{\theta}} \frac{\{(1+r_1)/r_2\}^{\lceil \theta \rceil}}{1 - (1+r_1)/r_2}. \end{aligned}$$

The right-hand side is

$$\frac{\sup_{|1-\mathbf{z}|=r_2} |f(\mathbf{z})|}{(1+r_1)^{\theta-\lceil \theta \rceil} r_2^{\lceil \theta \rceil} \{1 - (1+r_1)/r_2\}} \leq \left( \frac{be^{(r_2-1)^b}}{r_2} \right)^{\theta} \frac{(1+r_1)r_2}{r_2 - 1 - r_1}, \quad (5.11)$$

because

$$\begin{aligned} \sup_{|1-\mathbf{z}|=r_2} |f(\mathbf{z})| &= \sup_{|1-\mathbf{z}|=r_2} \left| \exp \left( -\theta \sum_{j=1}^b \frac{\mathbf{z}^j}{j} \right) \right| \\ &\leq \exp \left( \theta \sum_{j=1}^b \frac{(r_2-1)^j}{j} \right) \\ &\leq \exp \left( \theta \{ \log b + (r_2-1)^b \} \right) \\ &= (be^{(r_2-1)^b})^{\theta}, \end{aligned}$$

where we have used Lemma A.2 for the second inequality. Hence, it follows from the Cauchy inequality again that

$$\begin{aligned} |[x^{n-a}]h(x)| &\leq \frac{\sup_{|\mathbf{z}|=r_1} |h(\mathbf{z})|}{r_1^{n-a}} \\ &\leq \frac{1}{r_1^n} \left( \frac{be^{(r_2-1)^b}}{r_2} \right)^{\theta} \frac{1}{r_2 - 1 - r_1} \{r_1^a (1+r_1)r_2\}. \end{aligned}$$

This completes the proof.

Let us provide some remarks on Proposition 5.1.

**Remark 8.** Proposition 5.1 indicates that when  $\theta^2/n \rightarrow 0$  the components of  $(C_1^n, \dots, C_b^n)$  are asymptotically independent, and  $C_j^n$  asymptotically follows the Poisson distribution with mean  $\theta/j$  for  $j = 1, \dots, b$ . As a consequence, for any fixed  $b$ , if  $\theta \rightarrow c < \infty$  then

$$(C_1^n, \dots, C_b^n) \Rightarrow (Z_1, \dots, Z_b),$$

and if  $\theta \rightarrow \infty$  then

$$\frac{1}{\sqrt{\theta}} \left( C_1^n - \theta, \sqrt{2} \left( C_2^n - \frac{\theta}{2} \right), \dots, \sqrt{b} \left( C_b^n - \frac{\theta}{b} \right) \right) \Rightarrow N_b(0, I),$$

where  $N_b(0, I)$  is a  $b$ -dimensional standard normal variable with independent coordinates.

**Remark 9.** Proposition 5.9 below is stronger than Proposition 5.1, but the proof is included because some evaluations are different from the proof of Theorem 1 of Arratia, Barbour and Tavaré (1992) who used the Darboux lemma (see for instance Theorem of Knuth and Wilf (1989)), and because Lemmas 5.2 and 5.3 will be used in the proof of Theorem 5.8 below.

In Proposition 5.1,  $\theta^2/n \rightarrow 0$  is assumed. The following proposition shows that this assumption is necessary for the approximation of  $\{C_j^n\}_{j=1}^b$  by Poisson variables  $\{Z_j\}_{j=1}^b$ .

**Proposition 5.4.** In Case A, if  $\theta \geq 1$  for some  $n$ , then  $P(C_b^n = \mathbf{a}_b) \sim P(\mathbf{Z}_b = \mathbf{a}_b)$  for any  $\mathbf{a}_b$  with any fixed positive integer  $b$  only if  $\theta^2/n \rightarrow 0$ .

Proof. To prove the assertion, we see the case that  $b = 1$ . Let  $f(x) = \exp(-\theta x)$ , then we have  $f^{(k)}(x) = (-\theta)^k f(x)$ . From (5.3),  $g_{n-a} = [x^{n-a}](1-x)^{-\theta} f(x)$  equals

$$\frac{f(1)(\theta)_{n-a}}{(n-a)!} \sum_{k=0}^{\lceil \theta \rceil - 1} \frac{\theta^k (\theta - k)_{n-a}}{k! (\theta)_{n-a}} + [x^{n-a}]h(x).$$

Since

$$\begin{aligned} \frac{P(T_{1n} = n-a)}{P(T_{0n} = n)} &= \frac{\exp(-\theta \sum_{j=2}^n 1/j) g_{n-a}}{\exp(-\theta \sum_{j=1}^n 1/j) (\theta)_n / n!} = \frac{n!}{(\theta)_n} \frac{(\theta)_{n-a}}{(n-a)!} \sum_{k=0}^{\lceil \theta \rceil - 1} \frac{\theta^k (\theta - k)_{n-a}}{k! (\theta)_{n-a}} + o(1) \\ &\sim \sum_{k=0}^{\lceil \theta \rceil - 1} \frac{\theta^k (\theta - k)_{n-a}}{k! (\theta)_{n-a}} \end{aligned}$$

from the proof of Proposition 5.1, it is enough to show that

$$\sum_{k=0}^{\lceil \theta \rceil - 1} \frac{\theta^k (\theta - k)_{n-a}}{k! (\theta)_{n-a}} \rightarrow 1$$

only if  $\theta^2/n \rightarrow 0$ .

Since  $\theta$  is assumed not to decrease as  $n$  increases, we study the following three cases: (i)  $\theta \geq 2$  for some  $n$ ; (ii)  $\theta < 2$  for all  $n$  and  $\theta > 1$  for some  $n$ ; (iii)  $\theta \leq 1$  for all  $n$ . First, consider (i). Let  $n$  be a positive integer such that  $\theta \geq 2$ . Then, it holds that

$$\begin{aligned}
\sum_{k=0}^{\lceil\theta\rceil-1} \frac{\theta^k}{k!} \frac{(\theta-k)_{n-a}}{(\theta)_{n-a}} &\geq \sum_{k=0}^{\lceil\theta\rceil-2} \frac{\theta^k}{k!} \frac{(\theta-k)_{n-a}}{(\theta)_{n-a}} \\
&\geq \sum_{k=0}^{\lceil\theta\rceil-2} \frac{\theta^k}{k!} \frac{(\lceil\theta\rceil-1-k)_{n-a}}{(\lceil\theta\rceil-1)_{n-a}} \\
&= \sum_{k=0}^{\lceil\theta\rceil-2} \frac{\theta^k}{k!} \frac{(\lceil\theta\rceil-2)!}{(\lceil\theta\rceil-k-2)!} \frac{(\lceil\theta\rceil-2-k+n-a)!}{(\lceil\theta\rceil-2+n-a)!} \\
&= \sum_{k=0}^{\lceil\theta\rceil-2} \binom{\lceil\theta\rceil-2}{k} \theta^k \frac{1}{(\lceil\theta\rceil-2-k+1+n-a) \cdots (\lceil\theta\rceil-2+n-a)} \\
&= \sum_{k=0}^{\lceil\theta\rceil-2} \binom{\lceil\theta\rceil-2}{k} \left(\frac{\theta}{n}\right)^k \frac{1}{\left(1 + \frac{\lceil\theta\rceil-2-k+1-a}{n}\right) \cdots \left(1 + \frac{\lceil\theta\rceil-2-a}{n}\right)} \\
&\geq \sum_{k=0}^{\lceil\theta\rceil-2} \binom{\lceil\theta\rceil-2}{k} \left\{ \frac{\theta}{n \left(1 + \frac{\lceil\theta\rceil-2-a}{n}\right)} \right\}^k,
\end{aligned}$$

where we have used Lemma A.3 for the second inequality. From the binomial theorem, the right-hand side is equal to

$$\left\{ 1 + \frac{\theta}{n \left(1 + \frac{\lceil\theta\rceil-2-a}{n}\right)} \right\}^{\lceil\theta\rceil-2} = \left[ \left\{ 1 + \frac{\theta}{n \left(1 + \frac{\lceil\theta\rceil-2-a}{n}\right)} \right\}^{n/\theta} \right]^{\theta(\lceil\theta\rceil-2)/n}.$$

The above display is not less than 1 and converges to 1 only if  $\theta^2/n \rightarrow 0$ . Second, consider (ii). Let  $n$  be a positive integer such that  $\theta > 1$ . Then, it holds that

$$\sum_{k=0}^{\lceil\theta\rceil-1} \frac{\theta^k}{k!} \frac{(\theta-k)_{n-a}}{(\theta)_{n-a}} = 1 + \frac{\theta(\theta-1)}{\theta+n-a-1},$$

which converges to 1 only if  $\theta^2/n \rightarrow 0$ . Finally, consider (iii). Let  $n$  be a positive integer such that  $\theta = 1$ . Then, it holds that

$$\sum_{k=0}^{\lceil\theta\rceil-1} \frac{\theta^k}{k!} \frac{(\theta-k)_{n-a}}{(\theta)_{n-a}} = 1.$$

This completes the proof.

Thence, we have the following corollary to Propositions 5.1 and 5.4.

**Corollary 5.5.** *In Case A, if  $\theta \geq 1$  for some  $n$ , then  $P(\mathbf{C}_b^n = \mathbf{a}_b) \sim P(\mathbf{Z}_b = \mathbf{a}_b)$  for any  $\mathbf{a}_b$  with any fixed positive integer  $b$  if and only if  $\theta^2/n \rightarrow 0$ .*

Subsequently, let us derive the result corresponding to (2.15) following a similar programme to Arratia, Stark and Tavaré (1995). It follows from (2.8) that

$$d_b(n) = \sum_{a=0}^{\infty} \mathbb{P}(T_{0b} = a) \left( 1 - \frac{\mathbb{P}(T_{bn} = n - a)}{\mathbb{P}(T_{0n} = n)} \right)^+,$$

see (50) of Arratia, Stark and Tavaré (1995). Firstly, via the large deviation inequality, we see that  $d_b(n)$  is approximated by

$$\sum_{a=0}^{\lfloor J_n \rfloor} \mathbb{P}(T_{0b} = a) \left( 1 - \frac{\mathbb{P}(T_{bn} = n - a)}{\mathbb{P}(T_{0n} = n)} \right)^+$$

with

$$J_n = \min(b\theta \log n, b^{2/3}(\theta n)^{1/3}).$$

From the definition, if  $1 \leq b \leq n\theta^{-2}(\log n)^{-3}$  then  $J_n = b\theta \log n$  and otherwise  $J_n = b^{2/3}(\theta n)^{1/3} = b(\theta n/b)^{1/3}$ . In contrast to Arratia, Stark and Tavaré (1995),  $J_n$  includes  $\theta$  since we consider  $\theta \rightarrow \infty$ , but a similar treatment perform well.

**Lemma 5.6.** *In Case A, with  $b = o(n/\theta^2)$ , it holds that*

$$\sum_{a > J_n} \mathbb{P}(T_{0b} = a) \left( 1 - \frac{\mathbb{P}(T_{bn} = n - a)}{\mathbb{P}(T_{0n} = n)} \right)^+ \leq \mathbb{P}(T_{0b} > J_n) = o\left(\left(\frac{b}{n}\right)^k\right)$$

for any positive  $k$ .

Proof. The first inequality is obvious, so we see the latter one. From Lemma 8 of Arratia, Stark and Tavaré (1995), for any  $b \geq 1, w > 0$ , it holds that

$$\log \mathbb{P}(T_{0b} \geq bw) \leq \log(\theta e/w)^w. \quad (5.12)$$

If  $1 \leq b \leq n\theta^{-2}(\log n)^{-3}$  then, by putting  $w = \theta \log n$ , the right-hand side of (5.12) is

$$(\theta \log n)(1 - \log \log n) \sim -\theta(\log \log n) \log n$$

which tends to minus infinity faster than  $-k \log n$  for any positive  $k$ . If  $b \geq n\theta^{-2}(\log n)^{-3}$  then, by putting  $w = (\theta n/b)^{1/3}$ , the right-hand side of (5.12) is

$$\left(\frac{\theta n}{b}\right)^{1/3} \left(1 - \frac{1}{3} \log\left(\frac{n}{b}\right)\right) \sim -\frac{1}{3} \left(\frac{\theta n}{b}\right)^{1/3} \log\left(\frac{n}{b}\right)$$

which tends to minus infinity faster than  $-k \log(n/b)$  for any positive  $k$ . This completes the proof.

The next lemma shows that  $(|1 - \theta|/n) \mathbb{E}[(T_{0b} - \theta b)^+ 1\{T_{0b} \leq J_n\}]$  is approximately  $(|1 - \theta|/n) \mathbb{E}[(T_{0b} - \theta b)^+]$ .

**Lemma 5.7.** *In Case A, if  $\theta^2/n \rightarrow 0$  then it holds that*

$$\frac{|1 - \theta|}{n} \mathbb{E}[|T_{0b} - \theta b| 1\{T_{0b} > J_n\}] = o\left(\left(\frac{b}{n}\right)^k\right)$$

for  $b = o(n/\theta^2)$  and for any positive  $k$ .

Proof. From the Schwartz inequality, it follows that

$$\begin{aligned}
& \frac{|1 - \theta|}{n} \mathbb{E}[|T_{0b} - b\theta| \mathbf{1}\{T_{0b} > J_n\}] \\
& \leq \left( \frac{|1 - \theta|^2}{n^2} \mathbb{E}[(T_{0b} - \theta b)^2] \mathbb{P}(T_{0b} > J_n) \right)^{1/2} \\
& \leq \left( \frac{|1 - \theta|^2}{n^2} \theta b^2 \mathbb{P}(T_{0b} > J_n) \right)^{1/2} \\
& \leq \left[ \left\{ \frac{(1 + \theta)^2 b}{n} \right\}^2 \mathbb{P}(T_{0b} > J_n) \right]^{1/2}
\end{aligned}$$

where we have used  $\mathbb{E}[(T_{0b} - \mathbb{E}[T_{0b}])^2] = \text{var}(T_{0b}) = \sum_{j=1}^b j^2(\theta/j) = \theta \sum_{j=1}^b j \leq \theta b^2$  for the second inequality. Lemma 5.6 yields that  $\mathbb{P}(T_{0b} > J_n) = o((b/n)^{2k})$  for any positive  $k$ . This completes the proof.

The following result is an extension of (2.15) to large  $\theta$  setup.

**Theorem 5.8.** *In Case A, if  $\theta \geq 1$  for some  $n$  and  $\theta^2/n \rightarrow 0$ , then*

$$d_b(n) = \frac{\theta - 1}{2n} \mathbb{E}[|T_{0b} - \theta b|] + o\left(\frac{b\theta^2}{n}\right) \quad (5.13)$$

for

$$b = o\left(\frac{n}{\theta^2 \log n}\right). \quad (5.14)$$

In addition, when  $\theta \rightarrow \infty$ , it holds that  $d_b(n) = o(b\theta^2/n)$ .

Proof. Let  $n$  be a positive integer such that  $\theta \geq 1$ . Since it follows from Lemma 5.6 that

$$d_b(n) = \sum_{a=0}^{\lfloor J_n \rfloor} \mathbb{P}(T_{0b} = a) \left( 1 - \frac{\mathbb{P}(T_{bn} = n - a)}{\mathbb{P}(T_{0n} = n)} \right)^+ + o\left(\left(\frac{b}{n}\right)^k\right)$$

for any positive  $k$ , we see the first term.

Let  $g_{n-a} = \exp(\theta \sum_{j=b+1}^n 1/j) \mathbb{P}(T_{bn} = n - a)$ . For  $a \leq \lfloor J_n \rfloor$ , we have

$$\begin{aligned}
g_{n-a} &= \frac{f(1)(\theta)_{n-a}}{(n-a)!} \left\{ 1 + \frac{b\theta(\theta-1)}{\theta+n-a} + O\left(\frac{b^2\theta^4}{n^2}\right) \right\} + [x^{n-a}]h(x) \\
&= \frac{f(1)(\theta)_{n-a}}{(n-a)!} \left\{ 1 + \frac{b\theta(\theta-1)}{n} \left( 1 + O\left(\frac{a+\theta}{n}\right) \right) + O\left(\frac{b^2\theta^4}{n^2}\right) \right\} + [x^{n-a}]h(x)
\end{aligned}$$

and the last term should be evaluated for  $a$  and  $b$  growing with  $n$ .

If  $b$  does not diverge, as it is seen in the proof of Proposition 5.1,  $[x^{n-a}]h(x) = o(1/n^2)$  since  $a/n \leq J_n/n \rightarrow 0$ .

Thence, we consider the case that  $b \rightarrow \infty$ . Using Lemma 5.3 with  $r = 1/b$ , we have

$$|[x^{n-a}]h(x)| \leq \frac{b}{(1 + c_1/b)^{n-a}} \left( \frac{be^{(1+c_2/b)^b}}{2 + c_2/b} \right)^\theta \frac{(2 + c_1/b)(2 + c_2/b)}{(c_2 - c_1)}.$$

Since  $(1 + c_1/b)^{n-a} \sim e^{(n-a)c_1/b}$  and since  $(1 + c_2/b)^b \sim e^{c_2}$ , the right hand side is asymptotically equal to

$$\begin{aligned} \frac{b^{\theta+1} A_1^\theta}{\exp((n-a)c_1/b)} A_2 &= A_2 \exp \left( (\theta+1) \log b + \theta \log A_1 - \frac{(n-a)c_1}{b} \right) \\ &= A_2 \exp \left( \frac{n}{b} \left\{ \frac{b(\theta+1) \log n \log b}{n \log n} + \frac{b\theta}{n} \log A_1 - \left(1 - \frac{a}{n}\right) c_1 \right\} \right), \end{aligned}$$

where  $A_1 = (\exp(e^{c_2}))/2$  and  $A_2 = 4/(c_2 - c_1)$ . From (5.14), the right-hand side is

$$A_2 \exp \left( -\frac{n}{b} (c_1 + o(1)) \right) = A_2 \exp \left( -\frac{n(c_1 + o(1))}{b \log n} \log n \right) = A_2 n^{-\frac{n(c_1 + o(1))}{b \log n}}, \quad (5.15)$$

where we have used  $a/n \leq J_n/n \rightarrow 0$ . The right-hand side is  $o(1/n^k)$  for any positive constant  $k$ . After all, we have  $||x^{n-a}]h(x)| = o(b/n^2)$  even when  $b \rightarrow \infty$ .

Now we have

$$g_{n-a} = \frac{f(1)(\theta)_{n-a}}{(n-a)!} \left\{ 1 + \frac{b\theta(\theta-1)}{n} \left( 1 + O\left(\frac{a+\theta}{n}\right) \right) + O\left(\frac{b^2\theta^4}{n^2}\right) \right\} + o\left(\frac{b}{n^2}\right).$$

This expansion,  $f(1) = \exp(-\sum_{j=1}^b 1/j)$ ,  $P(T_{bn} = n-a) = \exp(-\theta \sum_{j=b+1}^n 1/j) g_{n-a}$  and (5.5)–(5.7) yield that

$$\frac{P(T_{bn} = n-a)}{P(T_{0n} = n)} = \frac{n!(\theta)_{n-a}}{(n-a)!(\theta)_n} \left\{ 1 + \frac{b\theta(\theta-1)}{n} + o\left(\frac{b\theta^2}{n}\right) \right\} + o(n)o\left(\frac{b}{n^2}\right).$$

Since  $a\theta/n \leq aJ_n/n \rightarrow 0$  and  $a^2/(nb) \leq J_n^2/(nb) \rightarrow 0$  which follow from  $\theta/J_n \rightarrow 0$  and (5.14), the binomial expansion and Lemma A.1 yield that

$$\begin{aligned} \frac{(\theta)_{n-a}/(n-a)!}{(\theta)_n/n!} &= \frac{(n-a)^{\theta-1} \left\{ 1 + \frac{\theta(\theta-1)}{2(n-a)} + O\left(\frac{\theta^4}{n^2}\right) \right\} / \Gamma(\theta)}{n^{\theta-1} \left\{ 1 + \frac{\theta(\theta-1)}{2n} + O\left(\frac{\theta^4}{n^2}\right) \right\} / \Gamma(\theta)} \\ &= \left(1 - \frac{a}{n}\right)^{\theta-1} \frac{\left\{ 1 + \frac{\theta(\theta-1)}{2n} + O\left(\frac{\theta^2(a+\theta^2)}{n^2}\right) \right\}}{\left\{ 1 + \frac{\theta(\theta-1)}{2n} + O\left(\frac{\theta^4}{n^2}\right) \right\}} \\ &= \left\{ 1 - \frac{a(\theta-1)}{n} + O\left(\frac{a^2\theta^2}{n^2}\right) \right\} \left( 1 + O\left(\frac{\theta^2(a+\theta^2)}{n^2}\right) \right), \end{aligned}$$

and hence

$$\frac{n!(\theta)_{n-a}}{(n-a)!(\theta)_n} = 1 - \frac{(\theta-1)a}{n} + O\left(\frac{\theta^2(a^2+\theta^2)}{n^2}\right) = 1 - \frac{(\theta-1)a}{n} + o\left(\frac{b\theta^2}{n}\right).$$

Therefore, it holds that

$$\begin{aligned} \frac{P(T_{bn} = n-a)}{P(T_{0n} = n)} &= \left\{ 1 - \frac{(\theta-1)a}{n} + o\left(\frac{b\theta^2}{n}\right) \right\} \left\{ 1 + \frac{b\theta(\theta-1)}{n} + o\left(\frac{b\theta^2}{n}\right) \right\} + o\left(\frac{b}{n}\right) \\ &= 1 - \left\{ \frac{(\theta-1)a}{n} - \frac{b\theta(\theta-1)}{n} \right\} + o\left(\frac{b\theta^2}{n}\right). \end{aligned}$$

From what has already been proven, it holds that

$$\begin{aligned}
d_b(n) &= \sum_{a=0}^{\lfloor J_n \rfloor} \mathbb{P}(T_{0b} = a) \left( \frac{(\theta - 1)a}{n} - \frac{b\theta(\theta - 1)}{n} \right)^+ + o\left(\frac{b\theta^2}{n}\right) \\
&= \frac{1}{n} \sum_{a=0}^{\infty} \mathbb{P}(T_{0b} = a) ((\theta - 1)(a - b\theta))^+ 1\{a \leq J_n\} + o\left(\frac{b\theta^2}{n}\right) \\
&= \frac{1}{n} \mathbb{E} [((\theta - 1)(T_{0b} - b\theta))^+ 1\{T_{0b} \leq J_n\}] + o\left(\frac{b\theta^2}{n}\right) \\
&= \frac{1}{n} \mathbb{E} [((\theta - 1)(T_{0b} - b\theta))^+] + o\left(\frac{b\theta^2}{n}\right) \\
&= \frac{\theta - 1}{2n} \mathbb{E} [|T_{0b} - b\theta|] + o\left(\frac{b\theta^2}{n}\right)
\end{aligned}$$

where we have used Lemma 5.7 in the fourth equality and the relation  $\mathbb{E}[(T_{0b} - b\theta)^+] = \mathbb{E}[|T_{0b} - b\theta|]/2$ , which follows from  $\mathbb{E}[T_{0b} - b\theta] = 0$ , in the fifth equality.

Finally, consider the case that  $\theta \rightarrow \infty$ . It follows from the Jensen inequality that

$$\mathbb{E}[|T_{0b} - b\theta|] \leq \sqrt{\mathbb{E}[|T_{0b} - b\theta|^2]} = O(\theta^{1/2}b),$$

which implies  $d_b(n) = o(b\theta^2/n)$ . This completes the proof.

Next we discuss the Poisson process approximation via the Feller coupling (see Subsection 2.2). The following result follows directly from (2.14).

**Proposition 5.9.** *Suppose that  $\theta > 1$  for all  $n$  and  $\theta^2/n \rightarrow 0$ . In Case A,  $d_b^W(n) \rightarrow 0$  if, and only if,  $b = o(n/\theta^2)$ . In addition, when  $\theta \rightarrow \infty$ , it holds that*

$$d_b^W(n) \sim \frac{\theta^2}{n} b.$$

**Remark 10.** *When  $\theta \rightarrow \infty$ , Theorem 5.8 and Proposition 5.9 yield that*

$$d_b(n) = o\left(\frac{b\theta^2}{n}\right), \quad d_b^W(n) = \Theta\left(\frac{b\theta^2}{n}\right)$$

*with  $b = o(n/(\theta^2 \log n))$ , which shows that the asymptotic decay rates of  $d_b(n)$  and  $d_b^W(n)$  are different.*

Let  $S_n^k$  be the  $k$ -th shortest cycle length in a Ewens partition, that is,

$$S_n^k = \inf(j : C_1^m + \dots + C_j^m \geq k)$$

for  $k = 1, 2, \dots$  and  $S_n^k = \infty$  when there is no such  $j$ . See Section 2E of Arratia and Tavaré (1992). The following statement is a direct corollary to Theorem 5.8.

**Corollary 5.10.** *Let  $r$  be a positive integer such that  $r = o(n/\theta^2)$  and let  $\delta_r = \sum_{j=1}^r \theta/j$ . Under the assumption of Proposition 5.9,*

$$\mathbb{P}(S_n^k \leq r) \sim \sum_{x=0}^{k-1} e^{-\delta_r} \frac{\delta_r^x}{x!}.$$



Proof. Proposition 5.9 yields that

$$P(S_n^k \leq r) = P\left(\sum_{j=1}^r C_j^m < k\right) \sim P\left(\sum_{j=1}^r Z_j < k\right) = \sum_{x=0}^{k-1} e^{-\delta_r} \frac{\delta_r^x}{x!}.$$

This completes the proof.

**Remark 11.** Corollary 5.10 yields that, under the assumption of Proposition 5.9,  $P(S_n^1 = 1) \sim e^{-\theta}$ , so if  $\theta \rightarrow \infty$  then  $P(S_n^1 = 1) \rightarrow 1$  and if  $\theta \rightarrow c < \infty$  then  $P(S_n^1 = 1) \rightarrow e^{-c} < 1$ . Note that when the Pitman sampling formula is considered, the shortest cycle length converges to 1 in probability except the Ewens sampling formula (see Mano (2017)).

The uniform bound with respect to  $b$ , which gives an extension of (2.13), is given in the following proposition. Its applications to functional central limit theorems will be presented in the next section.

**Proposition 5.11.** In Case A,

$$d_n^W(n) = O(\theta \log(1 + \theta)).$$

Proof. By using the triangle inequality and (2.12), it holds that

$$d_n^W(n) \leq \sum_{j=1}^n E[|C_j^m - C_j^\infty|] \quad (5.16)$$

$$\begin{aligned} &\leq \sum_{j=1}^b E[|C_j^m - C_j^\infty|] + \sum_{j=b+1}^n E[C_j^m] + \sum_{j=b+1}^n E[C_j^\infty] \\ &\leq \frac{b\theta(\theta+1)}{\theta+n-b} + 1 + 2\theta \log\left(\frac{n}{b}\right), \end{aligned} \quad (5.17)$$

for any  $b = 1, 2, \dots, n$ , see the proof of Theorem 2 of Arratia, Barbour and Tavaré (1992). When  $\theta \rightarrow \infty$ , by setting  $b = \lfloor n/\theta \rfloor$ , the first and third terms in (5.17) are  $O(\theta)$  and  $O(\theta \log \theta)$ , respectively. Otherwise, by setting  $b = \lfloor n/2 \rfloor$  the result holds with the bound  $d_n^W(n) = O(1)$ . Hence,  $d_n^W(n) = O((\theta \log \theta) \vee 1) = O(\theta \log(1 + \theta))$ . This completes the proof.

## 5.2 Case C

The probability mass function (1.1) is obtained from the conditioning relation (2.8) with a sequence of independent Poisson variables with respective means  $\theta/j$ . We also get (1.1) from (2.8) with Poisson variables with respective means  $(\theta/j)(n/\theta)^j$  (see for instance Watterson (1974a)). In Case C, the following lemma shows that  $E[Z_j] = (\theta/j)(n/\theta)^j$  rather fit.

**Lemma 5.12.** In Case C,

$$E[C_j^m] \sim \frac{\theta}{j} \left(\frac{n}{\theta}\right)^j.$$

Therefore, for  $j = 2, 3, \dots$ , if  $\theta(n/\theta)^j \rightarrow 0$ , then

$$C_j^n \rightarrow^p 0. \quad (5.18)$$

Proof. It holds that

$$\mathbb{E}[C_j^n] = \frac{\theta}{j} \frac{n!}{(n-j)!} \frac{\Gamma(n+\theta-j)}{\Gamma(n+\theta)}$$

which is (2.18) of Watterson (1974a). Since the Stirling formula  $\Gamma(x) = \sqrt{2\pi}x^{x-1/2}e^{-x} + O(x^{x-3/2}/e^x)$  as  $x \rightarrow \infty$  yields that  $\Gamma(x-c)/\Gamma(x) \sim x^{-c}$  as  $x \rightarrow \infty$  for any  $c < x$ , it holds that

$$\mathbb{E}[C_j^n] \sim \frac{\theta}{j} n^j \frac{1}{(n+\theta)^j} \sim \frac{\theta}{j} \left(\frac{n}{\theta}\right)^j.$$

Hence, the result (5.18) follows from  $C_j^n \geq 0$ . This completes the proof.

**Remark 12.** Since

$$\frac{n!}{(n-j)!n^j} \leq \frac{n!}{(n-j)!(n-j+1)_j} = 1$$

and

$$\frac{\Gamma(n-j+\theta)\theta^j}{\Gamma(n+\theta)} = \frac{\theta^j}{(n-j+\theta)_j} \leq 1,$$

it holds that

$$\frac{\mathbb{E}[jC_j^n]}{\theta(n/\theta)^j} = \frac{n!}{(n-j)!n^j} \frac{\Gamma(n-j+\theta)\theta^j}{\Gamma(n+\theta)} \leq 1 \quad (5.19)$$

for  $j = 1, 2, \dots, n$ .

According to Lemma 5.12, it may be natural to consider that  $C_j^n$  and Poisson variable with mean  $(\theta/j)(n/\theta)^j$  are asymptotically similar, but Proposition 3.3 indicates that, except Case C3, an independent process approximation by Poisson variables with means  $(\theta/j)(n/\theta)^j$  seems difficult in the sense of the joint distribution. Actually, the following theorem shows that in Case C2 the linear relation  $n - (C_1^n + 2C_2^n) \Rightarrow 0$  between  $C_1^n$  and  $C_2^n$  asymptotically remains.

**Theorem 5.13.** (i) In Case C2,

$$(C_1^n - n, C_2^n) \Rightarrow (-2P_{c/2}, P_{c/2}),$$

where  $P_{c/2}$  is a Poisson variable with  $\mathbb{E}[P_{c/2}] = \lim_{n,\theta} \{n^2/(2\theta)\} = c/2$ , and  $\sum_{j=3}^n |C_j^n| \Rightarrow 0$ .

(ii) In Case C3,

$$\sum_{j=1}^n |C_j^n - n1\{j=1\}| \Rightarrow 0.$$

Proof. (i) It follows from (5.19) that

$$\mathbb{E} \left[ \sum_{j=3}^n jC_j^n \right] \leq \sum_{j=3}^n \theta \left(\frac{n}{\theta}\right)^j = \frac{n^3}{\theta^2} \frac{1 - (n/\theta)^{n-2}}{1 - n/\theta},$$

which implies that

$$n - (C_1^n + 2C_2^n) = \sum_{j=3}^n jC_j^n \rightarrow^p 0.$$

It yields that  $\sum_{j=3}^n |C_j^n| \rightarrow^p 0$  and so

$$K_n - (C_1^n + C_2^n) \rightarrow^p 0.$$

These displays yield that

$$K_n - n + C_2^n \rightarrow^p 0.$$

From this, we obtain

$$\begin{aligned} & K_n - n - \left( \theta \log \left( 1 + \frac{n}{\theta} \right) - n \right) + C_2^n + \left( \theta \log \left( 1 + \frac{n}{\theta} \right) - n \right) \\ &= K_n - \theta \log \left( 1 + \frac{n}{\theta} \right) + C_2^n - \left( n - \theta \log \left( 1 + \frac{n}{\theta} \right) \right) \\ &\rightarrow^p 0. \end{aligned}$$

We conclude from (2.5), which means  $K_n - \theta \log(1 + n/\theta) \Rightarrow c/2 - P_{c/2}$ , that

$$C_2^n - \left( n - \theta \log \left( 1 + \frac{n}{\theta} \right) \right) = C_2^n - \frac{n^2}{2\theta} + o(1) \Rightarrow P_{c/2} - \frac{c}{2},$$

hence that  $C_2^n \Rightarrow P_{c/2}$ . Moreover, from what has already been proven, we obtain  $n - C_1^n \Rightarrow 2P_{c/2}$ .

(ii) Since

$$\mathbb{E}[|n - C_1^n|] = \mathbb{E}[n - C_1^n] = n - \frac{n}{1 + (n-1)/\theta} = \frac{n^2}{\theta} + O\left(\frac{n^3}{\theta^2}\right),$$

Lemma 5.12 yields the result. This completes the proof.

Theorem 5.13 directly implies the following corollaries which represent properties of the shortest cycle length  $S_n = S_n^1$  and the longest cycle length  $L_n$  in a Ewens partition. These extreme sizes are of interest in the combinatorial context, see for instance Mano (2017).

**Corollary 5.14.** *In Case C2 or C3,*

$$\mathbb{P}(S_n = 1) \rightarrow 1.$$

Proof. In Case C3, the conclusion is obvious, so we see Case C2. It holds that

$$\begin{aligned} \mathbb{P}(S_n \geq 2) &= \mathbb{P}(C_1^n = 0) = \mathbb{P}(n - C_1^n \geq n) \\ &\sim \mathbb{P}(2P_{c/2} \geq n) = 1 - \sum_{k < n/2} e^{-\frac{c}{2}} \left(\frac{c}{2}\right)^k \frac{1}{k!} \\ &\rightarrow 0. \end{aligned}$$

This completes the proof.

**Corollary 5.15.** *In Case C2, for a positive integer  $r$*

$$P(L_n \leq r) \sim \begin{cases} e^{-c/2} & (r = 1) \\ 1. & (r \geq 2) \end{cases}$$

Proof. For  $r = 1$ , it holds that

$$P(L_n \leq 1) = P(C_1^n = n) \sim P(W = 0) = e^{-c/2}.$$

For  $r \geq 2$ , it holds that  $P(L_n \leq r) \sim 1$ . This completes the proof.

**Remark 13.** *Since the law of the singleton  $C_1^n$  in a Ewens partition is given by*

$$P(C_1^n = k) = \frac{\theta^k}{k!} \left\{ \sum_{j=0}^{n-k} (-1)^j \frac{\theta^j (n+1-k-j)_{k+j}}{j! (n+\theta-k-j)_{k+j}} \right\}$$

for  $k = 0, 1, \dots, n$ , Corollary 5.15 directly follows from

$$\begin{aligned} P(C_1^n = n) &= \frac{\theta^n}{(\theta)_n} = 1 \frac{1}{1+1/\theta} \cdots \frac{1}{1+(n-1)/\theta} = 1 - \frac{1}{\theta} \frac{n(n-1)}{2} + o(1) \\ &= e^{-n^2/(2\theta)} + o(1). \end{aligned}$$

## 6 Functional central limit theorems

As a corollary to Proposition 5.11, the functional central limit theorems which extend the results of Hansen (1990) and Tsukuda (2017b) follow. Before the results, as a corollary to Proposition 5.11, let us states the error bounds of Poisson process approximations in the sense of the expectation of the error in the supremum norm and in the  $L^2$  norm.

**Corollary 6.1.** *In Case A,*

$$E \left[ \sup_{u \in [0,1]} \left| \sum_{j=1}^{\lfloor n^u \rfloor} \frac{(C_j^n - C_j^\infty)}{\sqrt{\theta \log n}} \right| \right] = O \left( \sqrt{\frac{\theta}{\log n}} \log(1 + \theta) \right) \quad (6.1)$$

and

$$E \left[ \left[ \int_0^1 \left\{ \frac{\sum_{j=1}^{\lfloor n^u \rfloor} (C_j^n - C_j^\infty)}{\sqrt{\sum_{j=1}^{\lfloor n^u \rfloor} \theta/j}} \right\}^2 du \right]^{1/2} \right] = O \left( \sqrt{\frac{\theta \log \log n}{\log n}} \log(1 + \theta) \right). \quad (6.2)$$

Proof. The result (6.1) follows from

$$E \left[ \sup_{u \in [0,1]} \left| \sum_{j=1}^{\lfloor n^u \rfloor} \frac{(C_j^n - C_j^\infty)}{\sqrt{\theta \log n}} \right| \right] \leq E \left[ \sup_{u \in [0,1]} \sum_{j=1}^{\lfloor n^u \rfloor} \frac{|C_j^n - C_j^\infty|}{\sqrt{\theta \log n}} \right] = \frac{\sum_{j=1}^n E [|C_j^n - C_j^\infty|]}{\sqrt{\theta \log n}},$$

and Proposition 5.11. The result (6.2) follows from

$$\int_0^1 \left| \frac{\sum_{j=1}^{\lfloor n^u \rfloor} (C_j^n - C_j^\infty)}{\sqrt{\sum_{j=1}^{\lfloor n^u \rfloor} \theta/j}} \right|^2 du \leq \frac{2 \left( \frac{1}{\log 2} - \log \log 2 + \log \log n \right)}{\theta \log n} \left( \sum_{j=1}^n |C_j^n - C_j^\infty| \right)^2$$

(for this evaluation see the proof of Lemma 3.1 of Tsukuda (2017b)) and from Proposition 5.11. This completes the proof.

By using Corollary 6.1, we provide functional central limit theorems which slightly extend the preceding result in which  $\theta$  is assumed to be fixed.

**Proposition 6.2.** (i) In Case A, if

$$\frac{\theta}{\log n} (\log \theta)^2 \rightarrow 0, \quad (6.3)$$

then the random process  $X_n^1(\cdot)$  defined in (2.16) converges weakly to a standard Brownian motion  $(B(u))_{0 \leq u \leq 1}$  in  $D[0, 1]$ .

(ii) In Case A, if

$$\frac{\theta \log \log n}{\log n} (\log \theta)^2 \rightarrow 0, \quad (6.4)$$

then both of the random processes  $X_n^2(\cdot)$  and  $X_n^3(\cdot)$ , which are defined in (2.17) and (2.18) respectively, converge weakly to  $(B(u)/\sqrt{u})_{0 < u < 1}$  in  $L^2(0, 1)$ .

Proof. (i) From (6.1) and the assumption (6.3), it follows that

$$\sup_{u \in [0, 1]} \left| \sum_{j=1}^{\lfloor n^u \rfloor} \frac{(C_j^m - C_j^\infty)}{\sqrt{\theta \log n}} \right| \leq \sup_{u \in [0, 1]} \sum_{j=1}^{\lfloor n^u \rfloor} \frac{|C_j^m - C_j^\infty|}{\sqrt{\theta \log n}} = \sum_{j=1}^n \frac{|C_j^m - C_j^\infty|}{\sqrt{\theta \log n}} \xrightarrow{p} 0. \quad (6.5)$$

By using the functional central limit theorem for Poisson processes in  $D[0, 1]$ , the random process

$$\left( \frac{\sum_{j=1}^{\lfloor n^u \rfloor} C_j^\infty - \sum_{j=1}^{\lfloor n^u \rfloor} \theta/j}{\sqrt{\sum_{j=1}^n \theta/j}} \right)_{0 \leq u \leq 1}$$

converges weakly to a standard Brownian motion  $(B(u))_{0 \leq u \leq 1}$  in  $D[0, 1]$  (see the Proof of Theorem 5 of Arratia and Tavaré (1992)). Since

$$\sup_{u \in [0, 1]} \left| \sum_{j=1}^{\lfloor n^u \rfloor} \frac{\theta}{j} - u\theta \log n \right| = O(\theta),$$

the random process

$$\left( \frac{\sum_{j=1}^{\lfloor n^u \rfloor} C_j^\infty - u\theta \log n}{\sqrt{\theta \log n}} \right)_{0 \leq u \leq 1} \quad (6.6)$$

converges weakly to  $(B(u))_{0 \leq u \leq 1}$  in  $D[0, 1]$  because of the assumption (6.3). From (6.5) and the weak convergence of (6.6), Theorem 2.7 (iv) of van der Vaart (1998) yields the result.

(ii) First we argue  $X_n^2(\cdot)$ . From (6.2) and (6.4), it follows that

$$\int_0^1 \left| \frac{\sum_{j=1}^{\lfloor n^u \rfloor} (C_j^m - C_j^\infty)}{\sqrt{\sum_{j=1}^{\lfloor n^u \rfloor} \theta/j}} \right|^2 du \xrightarrow{p} 0.$$

It holds that

$$\left( \frac{\sum_{j=1}^{\lfloor n^u \rfloor} C_j^\infty - \sum_{j=1}^{\lfloor n^u \rfloor} \theta/j}{\sqrt{\sum_{j=1}^{\lfloor n^u \rfloor} \theta/j}} \right)_{0 < u < 1} \stackrel{=d}{=} \left( \frac{N^1(\sum_{j=1}^{\lfloor n^u \rfloor} \theta/j) - \sum_{j=1}^{\lfloor n^u \rfloor} \theta/j}{\sqrt{\sum_{j=1}^{\lfloor n^u \rfloor} \theta/j}} \right)_{0 < u < 1}, \quad (6.7)$$

where  $(N^1(t))_{t \geq 0}$  is a homogeneous Poisson process with unit intensity. Since

$$\sup_{u \in (0,1)} \frac{|\sum_{j=1}^{\lfloor n^u \rfloor} \theta/j - u\theta \log n|}{\theta \log n} \rightarrow 0$$

and the other hypotheses hold with  $\lambda = 1$ ,  $s_n(u) = \sum_{j=1}^{\lfloor n^u \rfloor} \theta/j$  and  $f(n) = \theta \log n$  (see Subsection 6.2 of Tsukuda (2017b)) Lemma A.4 in Appendix implies that (6.7) converges weakly to  $(B(u)/\sqrt{u})_{0 < u < 1}$  in  $L^2(0,1)$ . From what has been already proven, Theorem 2.7 (iv) of van der Vaart (1998) yields the result.

Next we argue  $X_n^3(\cdot)$ . Since

$$\begin{aligned} \int_0^{\frac{\varepsilon}{\log n}} \frac{(N^1(u\theta \log n) - u\theta \log n)^2}{u\theta \log n} du &\rightarrow^p 0, \\ \int_{\frac{\varepsilon}{\log n}}^1 \frac{\left\{ \sum_{j=1}^{\lfloor n^u \rfloor} (C_j^n - C_j^\infty) \right\}^2}{u\theta \log n} du &\rightarrow^p 0, \\ \int_{\frac{\varepsilon}{\log n}}^1 \frac{\left( \sum_{j=1}^{\lfloor n^u \rfloor} C_j^\infty - N^1(u\theta \log n) \right)^2}{u\theta \log n} du &\rightarrow^p 0, \end{aligned}$$

which follow from the almost same argument as the proof of Theorem 7.1 of Tsukuda (2017b) by the assumption (6.4), we have

$$\int_0^1 \left( X_n^3(u) - \frac{N^1(u\theta \log n) - u\theta \log n}{\sqrt{u\theta \log n}} \right)^2 du \rightarrow^p 0.$$

From Lemma A.4 in Appendix with  $\lambda = 1$ ,  $s_n(u) = u\theta \log n$  and  $f(n) = \theta \log n$ , it holds that the random process

$$\left( \frac{N^1(u\theta \log n) - u\theta \log n}{\sqrt{u\theta \log n}} \right)_{0 < u < 1}$$

converges weakly to  $(B(u)/\sqrt{u})_{0 < u < 1}$  in  $L^2(0,1)$ . Consequently, the desired result follows. This completes the proof.

**Remark 14.** It follows from Proposition 6.2 (i) that if (6.3) holds then (2.1) holds. But as it is stated in (2.5), the asymptotic normality of  $K_n$  holds for far larger  $\theta$ .

The following result, promised in Subsection 2.3, is an extension of Proposition 2.1.

**Proposition 6.3.** (i) In Case A, if (6.3) holds then the random process  $X_n^4(\cdot)$  defined in (2.19) converges weakly to a standard Brownian bridge  $(B^\circ(u))_{0 \leq u \leq 1}$  in  $D[0,1]$ .

(ii) In Case A, if (6.4) holds then the random process  $X_n^5(\cdot)$  defined in (2.20) converges weakly to  $(B^\circ(u)/\sqrt{u(1-u)})_{0 < u < 1}$  in  $L^2(0,1)$ .

Proof. (i) Since it holds that

$$\begin{aligned} X_n^4(u) &= \frac{\theta \log n}{K_n} \left( \frac{\sum_{j=1}^{\lfloor n^u \rfloor} C_j^n - u K_n}{\sqrt{\theta \log n}} \right) \\ &= \frac{\theta \log n}{K_n} \left\{ \frac{(1-u) \sum_{j=1}^{\lfloor n^u \rfloor} C_j^n - u \sum_{j=\lfloor n^u \rfloor+1}^n C_j^n}{\sqrt{\theta \log n}} \right\} \end{aligned}$$

for any  $u \in [0, 1]$ , it is sufficient to show

$$\frac{K_n}{\theta \log n} \rightarrow^p 1 \quad (6.8)$$

and

$$\left( \frac{(1-u) \sum_{j=1}^{\lfloor n^u \rfloor} C_j^n - u \sum_{j=\lfloor n^u \rfloor+1}^n C_j^n}{\sqrt{\theta \log n}} \right)_{0 \leq u \leq 1} \Rightarrow (B^\circ(u))_{0 \leq u \leq 1} \quad (6.9)$$

in  $D[0, 1]$ . Firstly, (6.8) holds because the assumption (6.3) yields that  $\log \theta / \log n \rightarrow 0$  and because it follows from (2.7) that

$$\frac{K_n}{\theta \log n} = \frac{K_n}{\theta \log(n/\theta) \left(1 + \frac{\log \theta}{\log(n/\theta)}\right)} \rightarrow^p 1.$$

Next, we show (6.9). Since it follows from Proposition 5.11 and from the assumption (6.3) (see (6.5)) that

$$\sup_{u \in [0, 1]} \left| \frac{\sum_{j=1}^{\lfloor n^u \rfloor} C_j^n - \sum_{j=1}^{\lfloor n^u \rfloor} C_j^\infty}{\sqrt{\theta \log n}} \right| \leq \frac{\sum_{j=1}^n |C_j^n - C_j^\infty|}{\sqrt{\theta \log n}} \rightarrow^p 0$$

and that

$$\sup_{u \in [0, 1]} \left| \frac{\sum_{j=\lfloor n^u \rfloor+1}^n C_j^n - \sum_{j=\lfloor n^u \rfloor+1}^n C_j^\infty}{\sqrt{\theta \log n}} \right| \leq \frac{\sum_{j=1}^n |C_j^n - C_j^\infty|}{\sqrt{\theta \log n}} \rightarrow^p 0,$$

the triangle inequality yields that

$$\sup_{u \in [0, 1]} \left| \frac{(1-u) \sum_{j=1}^{\lfloor n^u \rfloor} C_j^n - u \sum_{j=\lfloor n^u \rfloor+1}^n C_j^n}{\sqrt{\theta \log n}} - P_4^\circ(u) \right| \rightarrow^p 0$$

where

$$(P_4^\circ(u))_{0 \leq u \leq 1} = \left( \frac{\sum_{j=1}^{\lfloor n^u \rfloor} C_j^\infty - u \sum_{j=1}^n C_j^\infty}{\sqrt{\theta \log n}} \right)_{0 \leq u \leq 1}.$$

By using the functional central limit theorem for Poisson processes in  $D[0, 1]$ ,  $P_4^\circ(\cdot)$  converges weakly to  $(B^\circ(u))_{0 \leq u \leq 1}$  in  $D[0, 1]$  (see Arratia and Tavaré (1992)).

(ii) By the same reason as (i), it is sufficient to show (6.8) and

$$\begin{aligned} &\left( \frac{(1-u) \sum_{j=1}^{\lfloor n^u \rfloor} C_j^n - u \sum_{j=\lfloor n^u \rfloor+1}^n C_j^n}{\sqrt{u(1-u)\theta \log n}} \mathbf{1}_{\left\{ \frac{\varepsilon}{\log n} < u < 1 - \frac{\varepsilon}{\log n} \right\}} \right)_{0 < u < 1} \\ &\Rightarrow \left( \frac{B^\circ(u)}{\sqrt{u(1-u)}} \right)_{0 < u < 1} \end{aligned} \quad (6.10)$$

in  $L^2(0, 1)$ . Firstly, it holds that

$$\begin{aligned}
& \int_{\frac{\varepsilon}{\log n}}^{1-\frac{\varepsilon}{\log n}} \frac{(1-u)^2 \left\{ \sum_{j=1}^{\lfloor n^u \rfloor} (C_j^n - C_j^\infty) \right\}^2}{u(1-u)\theta \log n} du \\
& \leq \int_{\frac{\varepsilon}{\log n}}^1 \frac{\left( \sum_{j=1}^{\lfloor n^u \rfloor} |C_j^n - C_j^\infty| \right)^2}{u\theta \log n} du \\
& \leq \int_{\frac{\varepsilon}{\log n}}^1 \frac{\left( \sum_{j=1}^n |C_j^n - C_j^\infty| \right)^2}{u\theta \log n} du \\
& \leq \frac{\log \log n - \log \varepsilon}{\theta \log n} \left( \sum_{j=1}^n |C_j^n - C_j^\infty| \right)^2, \tag{6.11}
\end{aligned}$$

and that

$$\begin{aligned}
& \int_{\frac{\varepsilon}{\log n}}^{1-\frac{\varepsilon}{\log n}} \frac{u^2 \left\{ \sum_{j=\lfloor n^u \rfloor+1}^n (C_j^n - C_j^\infty) \right\}^2}{u(1-u)\theta \log n} du \\
& \leq \int_0^{1-\frac{\varepsilon}{\log n}} \frac{\left( \sum_{j=\lfloor n^u \rfloor+1}^n |C_j^n - C_j^\infty| \right)^2}{(1-u)\theta \log n} du \\
& \leq \frac{\log \log n - \log \varepsilon}{\theta \log n} \left( \sum_{j=1}^n |C_j^n - C_j^\infty| \right)^2. \tag{6.12}
\end{aligned}$$

These right-hand sides of (6.11) and (6.12) converge to 0 in probability because the expectations of their square root converge to 0 by the assumption (6.4). Secondly, by letting  $(N^1(t))_{t \geq 0}$  be a homogeneous Poisson process with unit intensity, it follows from

$$\begin{aligned}
\mathbb{E}[(N^1(u\theta \log n) - uN^1(\theta \log n))^2] &= \text{var}(N^1(u\theta \log n) - uN^1(\theta \log n)) \\
&= (1-u)^2 u\theta \log n + u^2(1-u)\theta \log n \\
&= u(1-u)\theta \log n
\end{aligned}$$

that

$$\int_0^{\frac{\varepsilon}{\log n}} \frac{(N^1(u\theta \log n) - uN^1(\theta \log n))^2}{u(1-u)\theta \log n} du \xrightarrow{p} 0,$$

and that

$$\int_{1-\frac{\varepsilon}{\log n}}^1 \frac{(N^1(u\theta \log n) - uN^1(\theta \log n))^2}{u(1-u)\theta \log n} du \xrightarrow{p} 0.$$



Thirdly, it holds that

$$\begin{aligned}
& \int_{\frac{\varepsilon}{\log n}}^{1-\frac{\varepsilon}{\log n}} \frac{\left\{ \sum_{j=1}^{\lfloor n^u \rfloor} C_j^\infty - u \sum_{j=1}^n C_j^\infty - (N^1(u\theta \log n) - uN^1(\theta \log n)) \right\}^2}{u(1-u)\theta \log n} du \\
&= \int_{\frac{\varepsilon}{\log n}}^{1-\frac{\varepsilon}{\log n}} \frac{1}{u(1-u)\theta \log n} \left[ (1-u) \left( \sum_{j=1}^{\lfloor n^u \rfloor} C_j^\infty - N^1(u\theta \log n) \right) \right. \\
&\quad \left. - u \left\{ \sum_{j=\lfloor n^u \rfloor+1}^n C_j^\infty - (N^1(\theta \log n) - N^1(u\theta \log n)) \right\} \right]^2 du \\
&\leq \int_{\frac{\varepsilon}{\log n}}^{1-\frac{\varepsilon}{\log n}} \frac{2(1-u)^2 \left( \sum_{j=1}^{\lfloor n^u \rfloor} C_j^\infty - N^1(u\theta \log n) \right)^2}{u(1-u)\theta \log n} du \\
&\quad + \int_{\frac{\varepsilon}{\log n}}^{1-\frac{\varepsilon}{\log n}} \frac{2u^2 \left\{ \sum_{j=\lfloor n^u \rfloor+1}^n C_j^\infty - (N^1(\theta \log n) - N^1(u\theta \log n)) \right\}^2}{u(1-u)\theta \log n} du \\
&\leq \int_{\frac{\varepsilon}{\log n}}^1 \frac{2 \left( \sum_{j=1}^{\lfloor n^u \rfloor} C_j^\infty - N^1(u\theta \log n) \right)^2}{u\theta \log n} du \\
&\quad + \int_0^{1-\frac{\varepsilon}{\log n}} \frac{2 \left\{ \sum_{j=\lfloor n^u \rfloor+1}^n C_j^\infty - (N^1(\theta \log n) - N^1(u\theta \log n)) \right\}^2}{(1-u)\theta \log n} du.
\end{aligned}$$

The distributions of the first term and second term in the right-hand side are equal to

$$\int_{\frac{\varepsilon}{\log n}}^1 \frac{2 \left( N^1(\theta(\sum_{j=1}^{\lfloor n^u \rfloor} 1/j - u \log n)) \right)^2}{u\theta \log n} du$$

and

$$\int_0^{1-\frac{\varepsilon}{\log n}} \frac{2 \left( N^1(\theta(\sum_{j=1}^n 1/j - \log n)) - N^1(\theta(\sum_{j=1}^{\lfloor n^u \rfloor} 1/j - u \log n)) \right)^2}{(1-u)\theta \log n} du,$$

respectively. Both of them converge to 0 in probability because their expectations tend to 0 from the assumption (6.4). Thus, the triangle inequality yields that

$$\int_0^1 |X_n^5(u) - P_5^\circ(u)|^2 du \rightarrow^p 0,$$

where

$$(P_5^\circ(u))_{0 < u < 1} = \left( \frac{N^1(\theta u \log n) - uN^1(\theta \log n)}{\sqrt{u(1-u)\theta \log n}} \right)_{0 < u < 1}.$$

Since

$$\begin{aligned}
(P_5^\circ(u))_{0 < u < 1} &= \left( \int_0^{\theta \log n} \frac{1\{t \leq u\theta \log n\} - u}{\sqrt{u(1-u)\theta \log n}} dN^1(t) \right)_{0 < u < 1} \\
&= \left( \int_0^{\theta \log n} \frac{1\{t \leq u\theta \log n\} - u}{\sqrt{u(1-u)\theta \log n}} (dN^1(t) - dt) \right)_{0 < u < 1},
\end{aligned}$$

Theorem 4 of Tsukuda (2016) yields that  $(P_5^\circ(u))_{0 < u < 1} \Rightarrow (B^\circ(u)/\sqrt{u})_{0 < u < 1}$  by setting  $H_s = 1$  with  $d = 1$ ,  $\lambda_s = 1$  and  $T = \theta \log n$ . Consequently, we have (6.10). This completes the proof.

## A Appendix. Auxiliary lemmas

**Lemma A.1.** *In Case A, if  $\theta^2/n \rightarrow 0$  then*

$$\Gamma(\theta) \frac{(\theta)_n}{n!} = n^{\theta-1} \left\{ 1 + \frac{\theta(\theta-1)}{2n} \right\} + O\left(n^{\theta-1} \frac{\theta^4}{n^2}\right). \quad (\text{A.1})$$

Proof. The left-hand side of (A.1) equals

$$\frac{\Gamma(n+\theta)}{n\Gamma(n)}.$$

By using the asymptotic series expansion  $\Gamma(x) = \sqrt{2\pi}e^{-x}x^{x-1/2}\{1 + 1/(12x)\} + O(e^{-x}x^{x-5/2})$  as  $x \rightarrow \infty$ , it holds that

$$\begin{aligned} \Gamma(n+\theta) &= \sqrt{2\pi}e^{-(n+\theta)}(n+\theta)^{n+\theta-1/2} \left\{ 1 + \frac{1}{12(n+\theta)} \right\} + O\left(e^{-(n+\theta)}(n+\theta)^{n+\theta-5/2}\right) \\ &= \sqrt{2\pi}e^{-(n+\theta)}n^{n+\theta-1/2} \left( 1 + \frac{\theta}{n} \right)^{n+\theta-1/2} \left( 1 + \frac{1}{12n} + O\left(\frac{\theta}{n^2}\right) \right) \end{aligned}$$

and that

$$\begin{aligned} n\Gamma(n) &= n\sqrt{2\pi}e^{-n}n^{n-1/2} \left( 1 + \frac{1}{12n} \right) + O\left(e^{-n}n^{n-5/2}\right) \\ &= \sqrt{2\pi}e^{-n}n^{n+1/2} \left( 1 + \frac{1}{12n} \right) \left( 1 + O\left(\frac{1}{n^2}\right) \right). \end{aligned}$$

Hence, the left-hand side of (A.1) is

$$\begin{aligned} &n^{\theta-1}e^{-\theta} \left( 1 + \frac{\theta}{n} \right)^n \left( 1 + \frac{\theta}{n} \right)^{\theta-1/2} \left( 1 + O\left(\frac{\theta}{n^2}\right) \right) \left( 1 + O\left(\frac{1}{n^2}\right) \right) \\ &= n^{\theta-1}e^{-\theta} \left( 1 + \frac{\theta}{n} \right)^n \left( 1 + \frac{\theta}{n} \right)^{\theta-1/2} \left( 1 + O\left(\frac{\theta}{n^2}\right) \right) \end{aligned} \quad (\text{A.2})$$

and, from the asymptotic expansion  $(1 + 1/x)^x = e(1 - 1/(2x) + O(x^{-2}))$  as  $x \rightarrow \infty$  it follows that

$$\left( 1 + \frac{\theta}{n} \right)^n = e^\theta \left( 1 - \frac{\theta}{2n} + O\left(\frac{\theta^2}{n^2}\right) \right)^\theta$$

and hence (A.2) is

$$\begin{aligned} &n^{\theta-1} \left( 1 - \frac{\theta}{2n} + O\left(\frac{\theta^2}{n^2}\right) \right)^\theta \left( 1 + \frac{\theta}{n} \right)^\theta \left( 1 + \frac{\theta}{n} \right)^{-1/2} \left( 1 + O\left(\frac{\theta}{n^2}\right) \right) \\ &= n^{\theta-1} \left( 1 + \frac{\theta}{2n} + O\left(\frac{\theta^2}{n^2}\right) \right)^\theta \left( 1 + \frac{\theta}{n} \right)^{-1/2} \left( 1 + O\left(\frac{\theta}{n^2}\right) \right) \\ &= n^{\theta-1} \left( 1 + \frac{\theta^2}{2n} + O\left(\frac{\theta^4}{n^2}\right) \right) \left( 1 - \frac{\theta}{2n} + O\left(\frac{\theta^2}{n^2}\right) \right) \left( 1 + O\left(\frac{\theta}{n^2}\right) \right) \\ &= n^{\theta-1} \left\{ 1 + \frac{\theta(\theta-1)}{2n} + O\left(\frac{\theta^4}{n^2}\right) \right\} \left( 1 + O\left(\frac{\theta}{n^2}\right) \right) \\ &= n^{\theta-1} \left\{ 1 + \frac{\theta(\theta-1)}{2n} + O\left(\frac{\theta^4}{n^2}\right) \right\}. \end{aligned}$$

This completes the proof.

**Lemma A.2.** *Let  $b$  be an positive integer. For  $a > 1$ , it holds that*

$$\sum_{j=1}^b \frac{a^j}{j} \leq \log b + a^b.$$

Proof. It holds that

$$\sum_{j=1}^b \frac{a^j}{j} = \sum_{j=1}^b \int_0^a t^{j-1} dt = \sum_{j=1}^b \left( \frac{1}{j} + \int_1^a t^{j-1} dt \right) \leq 1 + \log b + \int_1^a \frac{t^b - 1}{t - 1} dt.$$

As for the last term, it holds that

$$\begin{aligned} \int_1^a \frac{t^b - 1}{t - 1} dt &= \int_1^a \frac{\{1 + (t - 1)\}^b - 1}{t - 1} dt \\ &= \int_1^a \sum_{j=1}^b \binom{b}{j} (t - 1)^{j-1} dt = \int_1^a \sum_{j=1}^b \binom{b-1}{j-1} \frac{b}{j} (t - 1)^{j-1} dt \\ &\leq \int_1^a b \sum_{j=1}^b \binom{b-1}{j-1} (t - 1)^{j-1} dt = \int_1^a b t^{b-1} dt \\ &= a^b - 1. \end{aligned}$$

This completes the proof.

**Lemma A.3.** *For any  $a > 0$  and for any positive integer  $b$ ,  $(x - a)_b / (x)_b$  is increasing with respect to  $x > a$ .*

Proof. The proof is by induction on  $b$ . When  $b = 1$ ,  $(x - a)/x = 1 - a/x$  is increasing. Let  $x_1$  and  $x_2$  satisfy  $a < x_1 < x_2$ . If the conclusion of the lemma is true for  $b$ , then the conclusion is also true for  $b + 1$  because

$$\frac{(x_2 - a)_{b+1}}{(x_2)_{b+1}} = \frac{x_2 + b - a}{x_2 + b} \frac{(x_2 - a)_b}{(x_2)_b} > \frac{x_2 + b - a}{x_2 + b} \frac{(x_1 - a)_b}{(x_1)_b} > \frac{x_1 + b - a}{x_1 + b} \frac{(x_1 - a)_b}{(x_1)_b} = \frac{(x_1 - a)_{b+1}}{(x_1)_{b+1}}.$$

This completes the proof.

**Lemma A.4.** *Let  $(N_t)_{t \geq 0}$  be a homogeneous Poisson process with intensity  $\lambda > 0$  satisfying  $N_0 = 0$ . Define the non-decreasing function  $(n, u) \mapsto s_n(u)$  with respect to  $0 \leq u \leq 1$  and with respect to  $n = 1, 2, \dots$  which satisfies  $\inf_{u \in (\tau, 1)} s_n(u) > 0$  for all  $0 < \tau < 1$ ,*

$$\lim_{n \rightarrow \infty} \left( \frac{\sup_{u \in (0, 1)} |s_n(u) - u f(n)|}{f(n)} \right) = 0 \quad (\text{A.3})$$

with  $n \mapsto f(n)$  an increasing function of  $n$  satisfying  $\lim_{n \rightarrow \infty} f(n) = \infty$ , and

$$\lim_{n \rightarrow \infty} \left\{ \int_0^1 \frac{du}{(s_n(u))^\delta} \right\} = 0$$

for some  $\delta > 0$ . Then, the random process

$$\left( \frac{N_{s_n(u)} - \lambda s_n(u)}{\sqrt{\lambda s_n(u)}} \right)_{0 < u < 1}$$

converges weakly to a Gaussian process  $(B(u)/\sqrt{u})_{0 < u < 1}$  in  $L^2(0, 1)$  as  $n \rightarrow \infty$ .

**Remark 15.** Lemma A.4 is a slight generalization of Lemma 2.1 of Tsukuda (2017b). The only difference is condition (A.3), where corresponding condition (2.1) of Tsukuda (2017b) is the case that  $f(n) = K \log n$  with a constant  $K$ . To show Lemma A.4, the equation

$$\lim_{n \rightarrow \infty} \left( \frac{s_n(u) \wedge s_n(v)}{\sqrt{s_n(u)s_n(v)}} \right) = \lim_{n \rightarrow \infty} \left( \frac{K((u \log n) \wedge (v \log n))}{\sqrt{K^2 \log n^u \log n^v}} \right) = \frac{u \wedge v}{\sqrt{uv}}$$

in the proof of Lemma 2.1 of Tsukuda (2017b) should be replaced by

$$\lim_{n \rightarrow \infty} \left( \frac{s_n(u) \wedge s_n(v)}{\sqrt{s_n(u)s_n(v)}} \right) = \lim_{n \rightarrow \infty} \left\{ \frac{(uf(n)) \wedge (vf(n))}{\sqrt{uf(n)vf(n)}} \right\} = \frac{u \wedge v}{\sqrt{uv}},$$

and the other part has no need to change.

## Acknowledgements

The author would like to his heartfelt gratitude to Professor Shuhei Mano for a lot of constructive comments.

## References

- Arratia, R., Barbour, A.D., Tavaré, S. (1992). Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Probab.* **2**, no. 3, 519–535.
- Arratia, R., Barbour, A.D., Tavaré, S. (2000). Limits of logarithmic combinatorial structures. *Ann. Probab.* **28**, no. 4, 1620–1644.
- Arratia, R., Barbour, A.D., Tavaré, S. (2016). Exploiting the Feller coupling for the Ewens sampling formula. *Statist. Sci.* **31**, no. 1, 27–29.
- Arratia, R., Stark, D., Tavaré, S. (1995). Total variation asymptotics for Poisson process approximations of logarithmic combinatorial assemblies. *Ann. Probab.* **23**, no. 3, 1347–1388.
- Arratia, R., Tavaré, S. (1992). Limit theorems for combinatorial structures via discrete process approximations. *Random Structures Algorithms* **3**, no. 3, 321–345.
- Barbour, A.D. (1992). Refined approximations for the Ewens sampling formula. *Random Structures Algorithms* **3**, no. 3, 2670–276.
- Barbour, A.D., Hall, P. (1984). On the rate of Poisson convergence. *Math. Proc. Cambridge Philos. Soc.* **95**, no. 3, 473–480.
- Crane, H. (2016). The ubiquitous Ewens sampling formula. *Statist. Sci.* **31**, no. 1, 1–19.

- DeLaurentis, J.M., Pittel, B.G. (1985). Random permutations and Brownian motion. *Pacific J. Math.* **119**, no. 2, 287–301.
- Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biology* **3**, 87–112
- Feng, S. (2007). Large deviations associated with Poisson–Dirichlet distribution and Ewens sampling formula. *Ann. Appl. Probab.* **17**, no. 5-6, 1570–1595.
- Feng, S. (2010). *The Poisson–Dirichlet distribution and related topics*. Models and asymptotic behaviors. Probability and its Applications (New York). Springer, Heidelberg.
- Feng, S. (2016). Diffusion processes and the Ewens sampling formula. *Statist. Sci.* **31**, no. 1, 20–22.
- Flajolet, P., Soria, M. (1990). Gaussian limiting distributions for the number of components in combinatorial structures. *J. Combin. Theory Ser. A* **53**, no. 2, 165–182.
- Hansen, J.C. (1990). A functional central limit theorem for the Ewens sampling formula. *J. Appl. Probab.* **27**, no. 1, 28–43.
- Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York.
- Knuth, D.E., Wilf, H.S. (1989). A short proof of Darboux’s lemma. *Appl. Math. Lett.* **2**, no. 2, 139–140.
- Mano, S. (2017). Extreme sizes in Gibbs-type exchangeable random partitions. *Ann. Inst. Statist. Math.* **69**, no. 1, 1–37.
- Tsukuda, K. (2016). A change detection procedure for an ergodic diffusion process. *Ann. Inst. Statist. Math.* doi:10.1007/s10463-016-0564-y.
- Tsukuda, K. (2017a). Estimating the large mutation parameter of the Ewens sampling formula. *J. Appl. Probab.* **54**, no.1, 42–54.
- Tsukuda, K. (2017b). Functional central limit theorems in  $L^2(0, 1)$  for logarithmic combinatorial assemblies. To appear in *Bernoulli*.
- van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, Cambridge.
- Watterson, G.A. (1974a). Models for the logarithmic species abundance distributions. *Theoret. Population Biology* **6**, 217–250.

- Watterson, G.A. (1974b). The sampling theory of selectively neutral alleles. *Advances in Appl. Probability* **6**, 463–488.
- Yamato, H. (2013). Edgeworth expansion for the number of distinct components associated with the Ewens sampling formula. *J. Japan Statist. Soc.* **43**, no. 1, 17–28.
- Yannors, N. (1991). Poisson approximation for random sums of Bernoulli random variables. *Statist. Probab. Lett.* **11**, no. 2, 161–165.